THE DESIGN OF AN ITEM BANK TO TEST READING IN ENGLISH AS A FOREIGN LANGUAGE

R.A. HILL

PH.D.

UNIVERSITY OF EDINBURGH

1988

TABLE OF CONTENTS

FIGURES

TABLES

ABSTRACT

An item bank is a set of test items calibrated on a single scale, independent of the sample used for calibration purposes. Key concepts which arise from this definition are those of 'dimensionality', 'ability', and 'difficulty', all of which must be investigated both at the pre- and post-test construction stages. A discussion of the dimensionality of reading in English as a foreign language (EFL) suggests that there is in fact a single dimension to a construct which is often analysed in terms of separately identifiable sub-skills. Testing 'ability' is best approached within a criterion-referenced framework, where rigorous attempts to ensure content validity are necessary at the test writing stage. Using criterion-referenced measurement as a philosophy of test construction helps in the analysis of what is and is not 'difficult' about EFL reading test items. Further investigation of difficulty highlights the important relationship between text and task, so that task construction should be a systematic procedure. Specific technical questions relating to item banks are discussed, in particular the use of the Rasch model is justified for the development of an EFL reading item bank. The principles discussed are applied to a practical testing situation, and the results analysed in terms of the key concepts. It is concluded that EFL reading is unidimensional and that the Rasch model is appropriate for use with data of this kind. An item bank for EFL reading in a particular context (Malaysian undergraduate testing) is derived from two separate anchor tests, which are shown to achieve similar results. Thus an item bank can be constructed, though it is debatable whether the methods of test analysis employed have any great advantage over traditional methods in situations where the population is relatively homogeneous or where test items are piloted on large representative samples of the population.

DECLARATION

I hereby declare that this thesis is my own work and that all sources have been duly acknowledged.

ACKNOWLEDGEMENTS

I wish to record my thanks to the Institute for Applied Language Studies of the University of Edinburgh, without whose scholarship scheme this thesis would not have been possible. I am also very grateful to Dr Clive Criper, the Director of the Institute, for much stimulating discussion in the early stages of this thesis.

I also wish to acknowledge my debt to Dr Alan Davies of the Department of Applied Linguistics in the University of Edinburgh for allowing my participation in the USM Placement Test project, data from which forms the basis for the practical part of this thesis.

The staff and students at USM Penang were extremely cooperative throughout the project, and I am happy to acknowledge the part they have played in developing the Placement Test.

It goes without saying that I alone am responsible for any shortcomings in the thesis.

CHAPTER 1

INTRODUCTION

Item banking is a potentially fertile research area for foreign language testing. Its chief virtue is its flexibility; in theory anyone who wishes to make measurements of achievement would be allowed access to a wide range of well-documented testing materials which could cover a variety of situations. A test user, knowing the general specification of items desired for a particular test, could select an item subset from the bank appropriate for his needs. Item banking allows the construction of cheap, secure and easily changeable large-scale tests, as well as individualised tests through computers.

The methods for constructing item banks have only been developed in the last 20 years or so, and controversy has surrounded their development. The present work looks into one small area of item banking - the design of an item bank to test reading in English as a foreign language - to examine the problems and difficulties that arise when the broad concept of 'item banking' is applied to one particular field.

Chapter 1 introduces some of the key concepts of item banking. Chapter 2 examines the dimensions of reading, while Chapter 3 focusses on the testing of reading and in particular on the underlying 'ability' that tests claim to measure. Chapter 4 analyses the idea of 'difficulty' in the context of reading tests. Chapter 5 explores some of the main ideas of the item bank model, and Chapter 6 describes the development of an item bank to test reading in English as a foreign language; a discussion of the results and their implications for future work is given in Chapter 7.

1.1. Definitions

1.1.1. Item

We take an item to be an individual question or unit which can be marked pass or fail. A test is therefore made up of a number of items. Another way of looking at an item is to consider it a test in itself, a miniature test usually consisting of a single problem (Lado 1961; 342). The score on a language item will then be a sample of the performance of students on a language problem.

1.1.2. Item Pool

Following Choppin (1976) and Childs (1978) we distinguish between an *item bank* and an *item pool*. An item pool is an unstructured collection of items. Although such an unstructured collection can be a valuable resource, one of the problems is that a test which is constructed by taking a subsample of items from such a collection cannot be used to compare performance on another test which was made up from a different sample of items designed to reflect a different curriculum, say.

An associated problem with this kind of item pool is that the very idea of sampling from a collection of items can be criticised as being conceptually unsound (Loevinger 1965). People can be sampled, but test items can not, since it makes little sense to talk of a *population* of test items (in spite of attempts to list test items exhaustively - e.g. Shoemaker 1976)

1.1.3. Item Bank

If a group of uncalibrated items is an item pool, then what distinguishes and defines an item bank is the calibration of individual items. The contribution of each item can be recorded separately in such a way that the characteristics of a test can be obtained by summing the characteristics of the individual items. Items can now be added or removed as freely as we wish without affecting the interpretation that can be put on the total test score.

In this case, an item bank is an "all-purpose measurement system" (Wood 1976) capable of meeting any testing requirement, group or individual.

1.2. Implications of the definitions

1.2.1. Item types

It follows from our definition of an item that we are not, on the whole, concerned with constructed responses to test items. Our preferred item type will be one where the testee selects the response from a determined choice (as in multiple choice questions for example). This should not, however, be taken to mean that constructed responses are ruled out by the idea of an item bank - essays are just as bankable as any other type of question (cf. Pollitt and Hutchinson 1987, e.g.). Rather, the criterion that an item be marked pass or fail tends to favour objective, selected response items. An incidental virtue of limiting ourselves to selected response items is that tests become entirely computer markable, thus saving time, money and resources which could be more usefully employed in other directions.

1.2.2. Item analysis

An item bank which contains substantial numbers of items which purport to measure the same dimension (which may be very loosely defined: 'Knowledge of English Vocabulary', for example) will be used in such a way that groups of items can be extracted from the bank and made to form an *ad hoc* test to provide more or less precise measurement of the trait in question. The bank then contains information as to what sub-group each item belongs to, but this hardly warrants the effort of bank development. As Choppin (1976;237) points out, if results on an item are to be interpreted, then one needs to know how difficult the item is and to what extent it discriminates between people of different ability. Unfortunately, the conventional measures of item difficulty and discrimination are 'sample bound', which means that they are heavily dependent on the nature of the sample of people who provided the data. Moreover, with traditional item analysis the interpretation of a test score is based on the test as a complete unit. The individual item is 'locked' into the test (Childs 1978;2) such that its contribution to any objective assessment is in its contribution to the total test score. If any of the items are removed or changed, the effect on the test as a whole is unknown. Hence traditional tests cannot be tampered with to provide a better reflection of a particular syllabus without altering the interpretation which can be made from the test score.

What is required is therefore a method of item analysis which allows precisely this possibility of selecting items with known statistical characteristics in order to construct tests which will not themselves alter the characteristics of the items.

1.2.3. Latent trait test models

The only statistical models that allow for the requirements outlined above are the so-called latent trait test models. In this group we include probabilistic models and item characteristic curve models. The mathematical differences between the models are not that great, but they originate from separate conceptions: the latent trait and ICC (or logistic) models arose from work being done in America, notably by Lord, to explain exactly how test items work in practice; probabilistic models arose from a concern that, especially in education, the response to a test item is never a matter of certainty and that there is always some chance, however remote, that a good candidate could fail an easy item; it was felt that a test model should reflect this reality. Such models have not been used traditionally because they are difficult to use from a mathematical point of view - in practical situations the problems involved can only be solved with the use of a large computer; furthermore, we have been quite happy to concentrate on whole tests rather than on individual items (for a variety of

reasons). This is now changing; we have the resources to solve the mathematical problems easily, and we should now begin to demand more of our tests.

1.2.3.1. The Rasch Model

The simplest of the latent trait models is the Rasch model. It is not, nor need be, universally accepted; but it shows in their starkest form the principles and the problems of latent trait measurement.

The simplest form of the Rasch model itself is as follows:

$$P = \frac{x}{1 + x}$$

where P = the probability of an individual answering a particular question right,
x = a function of the difference between a parameter representing the person's ability and the parameter representing the difficulty of the question.

The point about this function is that as x becomes very large, so P approaches 1, and as x becomes very small, so P approaches 0. Thus for all values of x the value of P lies somewhere between 0 and 1, which is the normal range for expressing probabilities. In the more usual, though more complex form, the equation looks like this:

$$P(+ \mid \nu,\iota) = \frac{\exp(\xi_\nu - \sigma_\iota)}{1 + \exp(\xi_\nu - \sigma_\iota)}$$

Where: $P(+ \mid \nu,\iota)$ = the probability that subject ν will score a correct response on item ι,
ξ_ν = the ability level of subject ν,
σ_ι = the difficulty of item ι.

The point about this function is that x in the first equation is now expressed in terms of two parameters (which we shall go on to term 'ability' and 'difficulty'), so if 'ability' is very much greater than 'difficulty' then the value of P approaches 1, while if 'ability' is very much lower than 'difficulty' then the value of P approaches 0. Ideally a

testee's ability should be at the same level as the difficulty of a test item, so the value of the combined expression will be 0, which means that the exponential function will be 1 (any number raised to the power 0 takes on the value 1), and so the value of P becomes 1 divided by 1+1, or 0.5. This encapsulates the insight that when a testee's ability matches the difficulty of the item he is taking then he should have a 50% chance of answering the item correctly. It also corresponds to, though is not identical with, the classical testing practice that a facility value of 0.5 represents the best chance of achieving a satisfactorily discriminating item.

1.2.3.2. 'Difficulty'

Although 'difficulty' is essentially the name given to one of the parameters in the above equation, the attempt to give it a real-life meaning is one of the key problems in item banking. It corresponds to traditional concerns with statements about what a test is a test of; the difference is that while this is a legitimate, though often unexamined, area of concern in traditional test theory, in all latent trait measurement it is an explicit problem.

As with traditional test theory, one assumption of the Rasch model is that the *order* of difficulty of items will remain constant from population to population even if the level of ability of the populations differs.

1.2.3.3. 'Ability'

Again, the term 'ability' is essentially the name given to one of the parameters in the latent trait equation, but it has real-life applications (it would not be very helpful if it did not). In traditional test theory 'difficulty' and 'ability' tend to be put together under more general considerations of 'test validity'; in latent trait applications, however, the two strands are teased out explicitly.

The problem of attaching semantic labels to what are basically syntactic definitions (cf. Lord and Novick 1968; 17) can lead to unnecessary confusion in the case of 'ability'. Just because we say we are measuring a subject's 'ability' does not mean that we are committed to using what would traditionally be called 'tests of ability' (or 'aptitude'). An achievement test can equally be a test of 'ability' in the latent trait sense.

On the other hand, tests of pure knowledge have to be excluded from latent trait tests. This is not such a problem for language testers, who have rarely in the past employed pure 'knowledge' questions, as it is for testers in certain other fields (e.g. O

grade history). But we should remember that we cannot ask questions such as "What is the English word for *sacerdos*?"

The question of how we might want to set about testing reading 'ability', and indeed of how we do so at present, is discussed in Chapter 3.

1.2.3.4. 'Sample-free' statistics

One of the chief claims made for the use of the Rasch model (or in fact any of the latent trait models) for testing purposes is that it enables us to obtain sample-free estimates of item difficulty and subject ability, unlike traditional test statistics which are heavily sample dependent. This will be more fully discussed in Chapter 5, but at this point it should be noted that 'sample-freeness' is a statistical artefact; as Wood (1976; 253) points out, empirical demonstration is not sufficient, and we should be wary of those who parade the sample-freeness of latent trait methods as if it were the new wonder ingredient.

The property of sample freeness derives from the use of regression analysis to estimate item parameters, a course of action which follows directly from the wish to calibrate items on an arbitrary collection of individuals rather than on a sample from a defined population, in which case a *correlational* method of estimation would be appropriate. We should constantly bear in mind that item-banking always implies this kind of regression analysis, and that if there are interaction effects between samples and items, so that items behave differently in different samples, whether because of cultural or sex bias or whatever, then invariance ceases to hold. The same applies to items which discriminate differentially across the ability range; if these items were to be calibrated on either high or low ability samples, the extrapolation of regressions would give a quite misleading picture of how the items behave.

The statistical derivation of this idea of 'sample-freeness' is given in Chapter 5, but we should always remember that this is one problem that latent trait analysis has not solved completely, even if it is more explicit about the matter than traditional analysis.

1.2.3.5. 'Item independence'

Another major constraint imposed by the use of the Rasch model is that items must be independent of each other. Again, this is a requirement of traditional testing models too, but it is not usually explicitly stated and is in any case less of a problem if we are considering whole test statistics rather than individual item statistics.

In practical terms, the requirement that items behave independently means that a reading comprehension passage with, say, 5 questions attached to it must be treated as a single 'item'.

1.2.3.6. Unidimensionality

A major requirement of an item bank is that it test only one dimension of the 'ability' under consideration. This is also true of traditional tests, though less explicitly stated. The problem of what actually constitutes a dimension is a difficult theoretical question which is addressed in Chapters 2 and 3; it is only partly answered by posterior statistical analyses.

One practical problem arising from this is alluded to by Childs (1978;11): the use of a single scale of achievement could encourage the teaching of a narrow area of the curriculum and the selection of items to represent this area. Similarly, the posterior identification of separate scales (and therefore separate dimensions) may lead either to a fragmentation of the trait when this is undesirable (e.g. separating 'ability to identify nouns in context' from 'ability to identify verbs in context') or to a looser definition of the trait when *this* is undesirable (e.g. suppressing the distinction between scientific reasoning and language comprehension); as Wood (1976;260-1) comments, fit to a statistical model is not necessarily informative about traits. The Rasch model leads to narrow trait definition because it imposes more stringent fit criteria than most models. The dilemma thus becomes whether to tighten up fit criteria and shrink domains or whether to keep to broad classes of items. The decision can only be made on pedagogic or logical grounds; whether and how we compartmentalize are important prior questions which we aim to tackle in Chapters 2 and 3.

1.2.4. Diagnostic testing

A concern with item banking implies a concern with diagnostic testing: not because one will always want to use an item bank for diagnostic purposes (though this should always be a possibility) but because the methods for categorising and labelling the items in a bank for retrieval purposes (as discussed in Chapter 5) share many of the ideas and methods of analysis of diagnostic and criterion-referenced testing. This will be discussed more fully in Chapter 3; at this juncture we simply set the diagnostic uses of item banking within a more general framework for the use of item banks.

There are at least two answers to the question of what a (Rasch-scaled) item bank

will look like (cf. Choppin 1979; 58-59); this is because an item bank for use by teachers within schools will differ quite considerably in its structure from the sort of item bank which has to be developed for the monitoring of standards on a large (national) scale. Both types of bank do, however, share the same logical base, and the results obtained from one should be able to be related to the results obtained from the other.

For monitoring purposes, a bank will need to have a very large number of test questions covering broad content areas and wide-ranging levels of difficulty. The collection of items is likely to be kept in some administrative centre and not published in full.

An item bank designed for teachers, on the other hand, would probably follow one of two designs: (a) the individual teachers might be offered a complete assessment service by some central agency. In this, teachers could specify the types of tests they needed and the agency would construct the tests and analyse the results (as is currently being tried out in South Australia and Tasmania). The NFER in Britain is also aiming for a system of this kind, but this is not yet available for comment. (b) The item bank could be published in the form of a book or pamphlet which contains all of the test questions together with details about what each measures in a test administrator's manual (as has been tried out for secondary school mathematics by Purushothaman 1975).

The point is that whichever of these latter two designs is adopted the bank will have to be structured to serve several alternative purposes which the teachers may demand. The bank should be capable of providing detailed diagnostic information about the attainment or otherwise of a particular skill or the mastery of a particular topic. It must therefore contain a large and varied selection of items catalogued accurately and in sufficient detail to permit the teacher to construct narrow but valid diagnostic tests. Other uses of the item bank are less problematical, but the minimum requirement is that some sort of diagnostic information must be available.

CHAPTER 2

THE DIMENSIONS OF EFL READING

2.1. Introduction

The usual requirement in test construction is that we first define or examine what it is that we claim to be testing: "All the normal criteria for evaluating the validity of test content should be applied from the earliest stages of test construction. The first principle therefore should be to as if the material to be included in items in the test is somehow related to the skill, construct or curriculum which the test is supposed to assess or measure." (Oller 1979;238). This is of particular importance in the case of item banking since the requirement of unidimensionality lays an explicit responsibility on us to discover what dimensions might be necessary. Heaton (1975;103) reminds us that since one of the chief concerns of the constructor of any test must always be to define the precise nature of what he is measuring, language testing can actually make a positive contribution to the development of reading skills here. Davies and Widdowson (1974; 170) similarly look for an answer to the question "What is it that we do when we fully comprehend written communication?" This is the question we address in this chapter.

To answer this major question we pose four subsidiary questions, which serve to guide the discussion:

1. What do we mean by 'reading' ?
2. Is reading in a foreign language different from reading in the native language?
3. Is reading in a foreign language different from other activities in the foreign language such as listening?
4. Is it possible to identify separate elements of reading in a foreign language?

2.2. Defining reading

2.2.1. Defining terms

The Shorter Oxford English Dictionary (1973 edition) has 17 definitions of the term 'reading'. Even in the professional literature prior assumptions are made as to the meaning of the term 'reading', so that what for one author might be a concern with the simple decoding of print, is for another a concern with more abstract 'higher-levels' of comprehension. The danger is that studies based on assumptions

that reading research should be all about decoding print will find their conclusions being used to justify statements which really belong to another area of research; thus Kolers's (1969) article "Reading is only incidentally visual" has had an influence far beyond its limited scope - originally intended as a study of the problem of visual perception it has on occasions (e.g. Coady 1979) been taken to justify a theory of FL reading which emphasises the overall perception of meaning rather than the creation of meaning by building up units of meaning within the text. This may well be a useful strategy for approaching FL reading but it needs to be justified within its own field, and not used unthinkingly as a conclusion from another 'universe of discourse'.

One major problem if one does not limit one's definition of reading to the ability to relate phonemes to graphemes is, as Alderson and Urquhart (1984a; xxvii) point out, that then any satisfactory definition of reading becomes all-embracing. Thus a note of despair is likely to be apparent: "Reading is as varied and adaptive an activity as perceiving, remembering or thinking, since in fact it includes all these activities" (Gibson and Levin 1975; 454), a sentiment echoed by La Berge and Samuels (1976; 576): "The complexity of the comprehension operation appears to be as enormous as that of thinking in general."

We can, perhaps identify two broad ways of defining reading; these are what Levin (1975; 125) calls the "sub-skills" of (a) decoding the writing system to its associated language and (b) using the code - the written version of the language - for the many uses to which reading may be put. Johnston (1983; 1), however, considers reading to be "any reader interaction with text"; thus comprehension is one aspect of reading, other aspects include decoding, scanning (e.g. phone directory use) and vocalizing the print on the page.

2.2.1.1. Implied opposites

One way of attempting to clarify the meaning of the term 'reading' is to try to identify the implied opposites that use of the term suggests. Thus Wiener and Cromer (1970) identify four basic polarities: first, Identification versus Comprehension. In other words, what behaviours define reading? "For those holding a single process view, identification [i.e. visual perception] can be considered a necessary antecedent to comprehension" (Wiener and Cromer 1970;137). But as the authors themselves point out, there may be several relationships involved rather than just this one. Considerations of this sort may lead one to debate whether, say, a foreigner 'reading' Italian words (i.e. just pronouncing them) can be considered to be reading in any useful sense. On one definition of reading this would certainly be the case. Such a

view of reading, indeed, underlies the 'Reading Universals Hypothesis' expounded (initially) by Goodman (1982): "It would seem that the reading processes will be much the same for all languages with minor variations to accomodate the specific characteristics of the orthography used and the grammatical structure of the language." (p.67)

Second, Acquisition versus Accomplished Reading. "In the acquisition of the reading skills, identification may be a necessary antecedent to comprehension ... But identification, which is essential in the acquisition phase for comprehension, may be irrelevant for the skilled reader ... who may go directly from the written forms to the meaning." (Wiener and Cromer 1970;138) On this view the final product of reading need not include components that went into its acquisition and therefore reading is rather like driving a car in that "in early learning there is much more cognitive behaviour associated with the sensory-motor behaviour while in the late phases operating a car is almost totally sensory-motor." (ib.;139) The question of which conscious 'skills' might be susceptible to training is an important one, as is the associated notion of identification of separable skills.

Third, Relative versus Absolute Criteria. "When absolute or ideal criteria are used, a good reader is typically specified as someone able to read a certain number of words at a given rate with some particular level of comprehension. Insofar as ideal criteria are arbitrary, standards can be designated which include differing proportions of the reading population." (Wiener and Cromer 1970; 139). The issue here is really that a relative definition of reading skill invokes criteria which specify, either implicitly or explicitly, some normative group, the implication being that the same kind or level of skill may be called 'good' or 'bad' reading depending on who is doing what and when. This is of particular relevance to L1 readers during their early years of schooling.

Fourth, Reading versus Language Skills. "The majority of research is less explicit [about the role of language in reading], even though comprehension implies the utilisation of meanings already available in some other (usually auditory) form." (Wiener and Cromer 1970;140).

There is thus the danger of ambiguity and confusion if we do not carefully define our terms; all definitions that focus on meaning or comprehension, for example, imply language as an antecedent but do not necessarily offer a basis for identifying poor reading as a reading difficulty rather than a language difficulty.

2.2.1.2. Reading as reasoning

There is a long tradition of viewing the essential concern of reading as the problem of reasoning, which includes such less wide-ranging ideas as 'inferencing'. Thorndike (1917) puts it as follows: "... reading is a very elaborate procedure, involving a weighing of each of many elements in a sentence, their organisation in the proper relationship one to another, the selection of certain of their connotations and the rejection of others, and the cooperation of many forces to determine final response. In fact ... the act of answering simple questions about a simple paragraph ... includes all the features characteristic of typical reasonings." (Thorndike 1917;323)

Reading research which follows this line has tended to degenerate into factor analysis (cf. e.g. Thorndike 1974), which, as we shall see, fails to resolve satisfactorily the theoretical issues. But Thorndike's (1917) proposals are concerned with quite specific features of reading and reasoning: pronouns, conjunctions, and prepositions especially which may have "meanings of many degrees of exactitude" (ib.;327). This concern led him to suggest that a very large percentage of the mistakes made in answer to simple questions about simple paragraphs were due to the "over-potency" of certain elements or the "under-potency" of others. Thus, for example, the question "What is the general topic of the paragraph?" produced answers as strange as : "The topic of the paragraph is one inch in."

For Thorndike (1917), then, understanding a paragraph is like solving a problem in mathematics: "The mind is assailed as it were by every word in the paragraph. It must select, repress, soften, emphasize, correlate and organize, all under the influence of the right mental set or purpose or demand." (op.cit.;329) A similar notion may be found in Widdowson (1978;63): "Reading is a kind of accomplishment whereby a discourse is created in the mind by means of a process of reasoning." There is a danger here, of course, of giving too much value to the subjective attitudes of the reader; the tension between 'reader's meaning' and 'textual meaning' is ever-present and difficult to resolve.

In the field of artificial intelligence, Charniak (1981) has drawn the distinction between 'problem solvers' and 'language comprehenders', the former being primarily concerned with deep inferences in narrow domains, the latter being more concerned with shallow inference in broader areas. However, as Charniak suggests (p.225), one's introspection suggests that the processes of language comprehension share much with problem solving and seem to rely on the same information.

He gives the example of comprehension test questions which often depend on

knowledge which would actually be useful if facing the 'problem' outlined in a problem-solving task used as a test passage. One is reminded also of Pask's (1976) distinction between 'comprehension learners' and 'potential holists' on the one hand and 'operation learners' and 'potential serialists' on the other, the former working by inference within the framework of global properties, the latter by piecing together their local knowledge of particulars. In both cases, reasoning is a crucial element in the route to understanding.

2.2.1.3. Summary

Reading is difficult to define - it can mean almost anything from simple letter recognition to complex reasoning. There is no single satisfactory definition. The rest of this chapter explores the various factors that need to be considered both in arriving at some sort of understanding of the term 'reading' and in investigating the activity itself.

2.3. Identifying comprehension

2.3.1. Models of the comprehension process

Models of reading are generally concerned with perceptual problems, but have an importance which goes beyond their immediate application. Models serve a guiding and an exploratory function; they can be descriptive or explanatory. A model of reading is essentially an abstraction from reality which is intended to order and simplify our view of that reality; it isolates and brings together ideas beliefs and knowledge about reading. The ordering involved in such model making will be selective, and models inevitably become *perceptual filters* shaping the individual's experience, influencing the kinds of questions asked, leading to particular expectations. Models can never really be 'objective'; Kennedy (1984; 46-53) has clearly demonstrated that a theory can actually cause us to observe the 'objective' world in a particular way, which may be at variance with what is 'really' the case. The implication of this for theories of reading is that failure to make a particular discrimination (for example between the letters *p* and *q*) does not mean that a child cannot see a difference between the two, but rather that the child does not yet have a *theory* about their difference; apparently perceptual failure is thus in fact a conceptual deficiency.

Goodacre (1979) distinguishes three types of reading model, based on the groups of people who need to define reading: the researcher's model (concerned basically with the cognitive aspects of reading), the test maker's model (concerned basically with sampling pupils' reading behaviour), and the teacher's model (where reading

tends to be thought of in terms of performance rather than as a thought- getting process, so that the situation can arise where a teacher can refer to a pupil as being 'good at reading' but 'poor at comprehension').

In what follows we shall focus on research techniques and limit the use of reading models to cognitive/ perceptual problems, and refer to other investigations into the nature of reading not as 'models' but rather as 'methods'.

The rationale for this approach stems from the observation that reading research has adopted one of two approaches: the *psychometric* approach and the *cognitive* approach (cf. Hewitt 1982; 10). The psychometric approach considers that reading comprehension consists of a number of sub-skills, that items can be written to test readers' operation of these sub-skills on texts, and that statistical analysis of subjects' responses to these items will reveal a number of factors comprising comprehension, or alternatively, will reveal that reading comprehension is a unitary skill.

The cognitive approach, on the other hand, is more interested in exploring how certain textual features, largely semantic features, interact with the reader's processing system. It is mainly interested in how readers process text and how the cognitive structures and processes which readers bring to the text interact with it to produce comprehension. It does this mainly by analysing readers' recall protocols in terms of how these are affected by certain variables such as textual features, cognitive structures or processes. This analysis then enables statements to be made about how readers process and comprehend texts. In addition, the cognitive approach is also interested in how the text is represented in memory and what variables affect the learning and recall of text.

2.3.1.1. Psycholinguistic models

The general criticisms of the use of psycholinguistic models to account for reading behaviour have been outlined by Hewitt (1982; 14-15). The first problem is that it cannot be assumed that what is recalled is necessarily comprehended and what is not recalled is not comprehended; it is quite possible to read a text and feel that one is comprehending it and yet be able to recall very little of the text. This amounts to a failure to answer directly the question 'What exactly is the relationship between 'reading comprehension' and 'memory for text'?'

A second criticism focusses on the nature of experimental and quasi- experimental designs. These generally attempt to focus upon or manipulate a small number of variables in a specific context using a small sample; this raises the fundamental

questions of (a) whether focussing on a small number of variables in isolation can ever produce understanding of a very complex behaviour; (b) whether the relationships found in highly controlled conditions will hold when a variety of confounding variables are introduced in a natural context; and (c) whether the effects discovered while using materials and tasks which are produced especially for the experimental situation will replicate with natural materials and tasks.

A third criticism is that very few studies have attempted to use their particular findings to develop a more general theory or model of reading comprehension processes (Kintsch and van Dijk 1978 are a notable exception here).

Finally, there is a group of criticisms aimed at cognitive approaches to understanding generally (including perception in the widest sense); these criticisms can be seen at their clearest in Ryle (1954). The argument is that perception should not be seen as the concluding stage of chain-processes, since it is not a state or process at all. The issues raised here would lead into complex philosophical analysis for which we are not equipped; we would note, however, that the development and use of computers tends to emphasize the analogy that the brain operates in the same way as some vastly complex computer, and that perception of all sorts is therefore to be seen as the concluding stage of a chain-process.

We now proceed to survey briefly the cognitive approach to the understanding of reading. Lovett (1981) has a four-fold classification of reading models which it will be useful to follow here.

2.3.1.2. Bottom-up models

Bottom-up approaches are typified by Gough (1972) and a model which he terms 'Merlin'. In this view, reading can be seen basically as the serial decoding of letters and as such the model pays no attention to the problem of integrating sentences and propositions, as Mitchell (1982) has pointed out. Moreover, the model lacks flexibility and higher order processes are isolated from lower order ones and hence cognition is effectively isolated from perception, whereas "most complex acts of information processing are accomplished through the interaction of higher and lower order processes ... and perception itself is an active, cognitively influenced operation." (Lovett 1981;3)

This type of model functions at an unconscious level, and as such it offers little basis for teachers or testers.

2.3.1.3. Top-down models

The second type of model that Lovett discusses embraces some of the more influential writers in the field. The models as a class can be seen as a reaction to the rigid reductionism of the linear models, exemplified in Gough; perhaps the best know exponents of this view are Goodman, Kolers and Smith, who all essentially subscribe to the view that reading is, in Goodman's phrase, *a psycholinguistic guessing game*.

What this means in practice is that reading is viewed as a continuous process of predicting, sampling, checking and revising. It is an almost exclusively 'top-down' approach which relies on and draws heavily from the analysis-by-synthesis principle of Halle and Stevens (1964 and 1967) to explain oral language comprehension and further articulated in Neisser's (1967) theory of cognition.

In this view the skilled reader is characterised as an active processor of textual information "circumventing laborious stages of letter-by-letter and even word-by-word perception ... The success of the reader's strategy will depend upon the natural redundancy of the language he samples and upon his own knowledge of linguistic constraints." (Lovett 1981;4)

This idea of redundancy surfaces particularly in Smith (1971): "Knowledge of redundancy constitutes a readily available, internalised source of information ... more meaning can be extracted and greater comprehension can be gained from the same number of visual features if syntactic and semantic sequential redundancy can be applied." (p.201)

As an integral part of this, the role of grammar becomes important: "The listener or reader uses grammar to comprehend the relationships that exist among the serially ordered elements of the sentence, to alter the organisation of his cognitive structure." (Smith 1971;194)

This notion of 'redundancy' and expectations created by the grammar of the language has been influential in a variety of ways in FL reading and pedagogy - notably in the use of cloze and in Oller's (1976) notion of an 'expectancy grammar'. This being the case, it is important to realise that the kind of psycholinguistic model upon which such ideas are based has in fact been subjected to a considerable amount of criticism; Lovett, for example, points out that these models as a class remain inadequate primarily for their failure to generate testable hypotheses: "There is considerable variability in the extent to which sampling theorists have been willing to operationalise the model ... there is as yet no convincing evidence that reading is, in

fact, a partial processing operation." (1981;4)

Similarly, Mitchell (1982) has pointed out that the main problem with Goodman's model is that it does not specify much about the reading process. Nor does it indicate how the various non-visual sources of information are drawn upon and used to modulate the formation of the perceptual image, and it does not say anything about how the system deals with the problem of graphic cues which are repeated in successive fixations (p.130). In addition, the lack of precision at different stages of processing means that it is difficult to determine exactly what claims the model makes about the process of reading.

It is true that the model is explicit about the fact that reading is a predictive process, that the reader samples from the print just enough to confirm his guess of what's coming next, but the evidence provides little support for these statements: "... a detailed review of the literature yielded no firm evidence that any of the processes that precede word recognition are influenced by the reader's anticipations." (Mitchell 1982;131)

Mitchell believes that the evidence for models of this type relies too much on experimental conditions using degraded material, where the processor (the reader, in this case) is forced to adopt a top-down approach because it is the only available strategy. This could offer a useful insight into the way the FL reader works, if we consider that he is rather in the position of someone working with impoverished stimuli. If this is the case, then the Goodman and Smith models may yet have something useful to say about FL reading, albeit for the wrong reasons (cf. Mitchell 1982;124).

2.3.1.4. Interactive models

The third type of model discussed by Lovett (1981) is exemplified in Rumelhart's (1977) *interactive* model, which can be seen as the revival of the notion that reading is a hypothesis testing operation. The model specifies that everything the reader needs in order to decode and understand print is organised into a series of six independent 'knowledge sources' viz. visual, featural, letter, letter cluster, lexical, syntactic semantic. The bi-directionality of the system is such that information furnished from any other knowledge source can affect current, past and subsequent contributions from any other knowledge source (Lovett 1981;5).

Mitchell (1982) points out that there have been various objections to this model, but that Rumelhart himself makes it clear that his main aim is to present a framework

for the development of models which is an alternative to the conventional serial flow-chart and places more emphasis on highly interactive parallel processing.

A study by Freebody and Anderson (1983) appears to lend weight to the interactive theory of reading. They point out that an interactive theory of reading assumes that reading involves many complementary levels of analysis and that a satisfactory understanding of a particular element in a text depends not only on an accurate identification of the words, but also on a knowledge of syntax, analysis of connections between this element and other parts of the text, and prior knowledge of the topic. An interactive theory of reading therefore gives rise to the "compensation hypothesis" (Freebody and Anderson 1983; 278): when one source of knowledge about the meaning of a text element is inoperative, other sources of knowledge may provide alternative ways of determining meaning.

In two experiments, Freebody and Anderson found that lack of connectives does not seriously damage comprehension because readers are usually able to make bridging inferences - reading merely becomes more effortful. On the other hand, vocabulary had a consistent, direct effect on performance, the theory being that many readers, on encountering a word they do not know, simply skip it, avoiding a drain on resources.

The conclusion (Freebody and Anderson 1983; 293) is that there is no support for the hypothesis that when one source of knowledge about the meaning of a text element is degraded, other sources of knowledge may compensate and provide alternate ways of determining meaning. Performance was lower when the passages contained difficult vocabulary, but it takes a "surprisingly high" proportion of difficult vocabulary to produce reliable decrements in comprehension measures. Thus it is probably a mistake to interpret the high correlations always seen between vocabulary tests and general tests of reading proficiency as indicating that word knowledge is of overriding instrumental importance in text comprehension. The more familiar version of a text was always better recalled; cohesion in the specific sense of linguistic ties, "simply is not very important in reading" (Freebody and Anderson 1983; 293)

2.3.1.5. Automaticity

The fourth and last type of model recognised by Lovett (1981) is that of LaBerge and Samuels (1974), which is based on the premise that all well-learned stimulus patterns can be encoded with or without attentional direction. They have contended that the development of automaticity (i.e. processing without attention) in all decoding processes is essential to fluent reading and that fluency is established only

when all levels of visual to semantic decoding proceed automatically and attention is thereby freed for continuous processing at the semantic level. However, the concern here seems to be with the lower-order perceptual skills and is really a theory about "the relationship between attention and the subprocesses of reading" (Mitchell 1982;134) and was not intended to be a comprehensive model of the reading process. A central claim of the model, as Mitchell points out, is that while fluent readers can carry out certain operations without attention, this is the outcome of considerable practice. LaBerge and Samuels argue (1976; 574) that from the point of view of a mature reader the process appears to be a unitary one. On the basis of this model, reading acquisition is viewed as a series of skills, regardless of how it appears to the fluent reader. Moreover, in its present form the model does not spell out higher-order linguistic operations such as parsing, predictive processing and contextual effects on comprehension.

2.3.1.6. Individual style

At this point it will be useful to present Mitchell's summary of what is involved in fluent reading:

> 1. The fluent reader recognises the majority of words in a text without pronouncing them implicitly or explicitly and without making use of contextual constraints;
>
> 2. Processing is carried out simultaneously at all different levels of the system - the reader's attention does not pass from one sub-process to another;
>
> 3. The processes that occur after word recognition make a significant contribution to the reading process as a whole;
>
> 4. The control and guidance of eye movements is an integral part of the reading process;
>
> 5. Reading is a flexible process. A fluent reader can suspend the more routine operations while he imagines a scene (say) or works out the implications of what he has just read. He can also skip words, sentences, or larger chunks of text if they do not seem essential to his immediate purpose.

(Mitchell 1982;136)

The first point that Mitchell makes here may seem a little surprising - that contextual constraints do not aid word recognition - but his review of the experimental evidence pointed strongly to the fact that readers make use of contextual information when the visual quality of the reading material is poor and that

word recognition is not guided or influenced in any way by the contextual information under normal conditions, but that the identification process runs its course and produces a decision which is then checked against the earlier material.

In this connection, and especially from the point of view of the FL reader, it is useful to recall Spiro's (1980) discussion of the aetiology of an individual's reading style. He isolates five causes of possible over- or under-reliance on text: Firstly, local or general schema unavailability may result in a text-based reading style.

Secondly, skills are not to be considered perfectly determinate of styles or vice versa, so that slow word identification, for example, can lead either to perseverance with decoding, thus creating a 'bottleneck' (Perfetti and Lesgold 1978) in higher order comprehension processes, or to an escape from this unpleasantness by compensating with reliance on top-down processing.

Thirdly, readers' own misconceptions about the reading process may create difficulties - they may think that reading is a bottom-up process and that top-down, extra-textual activities are inappropriate. In connection with this point in FL reading, Van Parreren and Schoutten-Van Parreren (1982) have shown that starting on the lexical level is 'dangerous' because most subjects do not possess the morphological knowledge required, but that in spite of the fact that hypotheses of this kind are risky and often unreliable, most subjects do not feel that. They act as if their hypotheses are quite firmly grounded and seem to have no difficulty in distorting and forcefully adapting the context to fit them. Hypotheses based on the context only are, on the other hand, mostly less trusted (see Van Parreren and Schoutten- Van Parreren 1982;239). This would account for the well-attested observation that FL readers often tend to focus and concentrate on the word rather than the sentence.

Fourthly, this aspect of an individual's reading style concerns general cognitive processing styles, which may dictate discourse processing styles.

Finally, there are breakdowns which create the appearance of over- reliance on the text. (For full details see Spiro 1980;263-4)

2.3.1.7. A developmental model

This final model is Huey's developmental model of reading as analysed by Sticht (1972). This analysis is of importance for the more general issue it raises of language as opposed to reading comprehension.

Sticht's analysis begins with Huey's statement: "The child comes to his first reader

with his habits of spoken language fairly well formed, and those habits grow more deeply set with every year. His meanings inhere in this spoken language and belong but secondarily to the printed symbols ..." From this, Sticht asserts that reading presupposes and is built upon a foundation of language ability. A simple extension of this model is that language ability presupposes and is built upon a base of pre-literate, pre-linguistic, perceptual/cognitive, adaptive, capabilities collectively referred to as 'intelligence' (Sticht 1972;291)

Thus, in Sticht's words, the model asserts a hierarchical developmental relationship among intelligence, listening (language) and reading such that language comprehension by reading depends upon and in fact encompasses the prior capability to comprehend language by listening. The latter in turn requires some 'core' intellectual capabilities for language to develop.

A consequence of this model would seem to be that, with mature readers, a measure of language comprehension by reading must simultaneously be a measure of intelligence and a measure of the ability to comprehend language (usually by listening), and a measure of the ability to read for comprehension.

What Sticht in fact found was a kind of implicational scale in which tests of 'listening' (or language - he uses the two more or less interchangeably, as may be justified in the L1 setting) include measurement of intelligence and tests of readinginclude both intelligence and listening.

One of the implications, both of Huey's model and of Sticht's findings, must be that there do not exist two kinds of language comprehension, one for reading and one for listening; "rather there is only one, wholistic ability to comprehend by language, and one should be able to comprehend equally well by listening or by reading, if one has been taught to decode well and other task variables are equalised." (Sticht 1972;293)

The importance of this is far-reaching; it helps justify the use of psycholinguistic findings related to speech perception in the field of written language decoding - something that Mitchell mentions but does not really elaborate upon. In fact, many psycholinguistic analyses of written language (and Mitchell is only one) are in fact analyses of language perception, evidence for which often, but by no means always, happens to have come through the written as opposed to the spoken mode. Moreover the whole of schemata-based approaches to comprehension depend upon this symbiotic relationship between reading and listening, where the question "What is reading?" usually means "What is language comprehension (whether written or

spoken)?" In fact, once one goes beyond the pure 'decoding of print' stage it is difficult to see how it could be otherwise.

Sticht concludes by pointing out that inasmuch as reading and listening both represent modes of comprehension by language, the major factor of concern is comprehension *by language*, rather than comprehension by reading or comprehension by listening. "Furthermore, it is to be desired and expected that with readers beyond the learning-to-decode reading stage, learning by listening and learning by reading should be highly correlated, as should these factors with other language (verbal) tests ..." (Sticht 1972;295) He also cites evidence that skills gained by listening could be transferred to reading test performance.

2.3.1.8. Hyperlexia

One curious phenomenon discovered by reading research needs to be mentioned here, because although it is a phenomenon found in deficient L1 readers, the description of this phenomenon fits remarkably well the observed behaviour of L2 readers. It is also presented here to demonstrate that interpretations of reading behaviour exclusively in terms of 'bottom up' or 'top down' processing can be misleading, in that over-reliance on one strategy may be a result of inadequate use of the other strategy. We refer to what Healy (1982) calls the "enigma" of hyperlexia.

There exists a group of children who exhibit advanced word-calling skills in the (reported) absence of meaningful comprehension of material read. Hyperlexia is defined as the presence of advanced word-calling abilities accompanied by marked deficits in comprehension. More than half of the children studied by Healy had been diagnosed as mentally retarded or severely deviant in development, though in fact their intellectual function ranged from mentally defective to above average. Expressive language skills were notably lacking. Social skills were also poorly developed; the children were characterized as "inflexible", "literal", and unable to understand other people. Disordered peer relations were common to all.

In these children, reading has assumed importance to the child as an activity in and of itself, rather than as one to obtain meaning. They have extreme difficulty with tasks requiring meaningful organization of verbal structures. When requested to give the meaning of common words the children are unable to formulate definitions beyond a single-word associative level (e.g "door" - "close"). Overall, reading, like their spontaneous speech, is intoned with a compulsive and stereotyped quality, though the reading might sound as if comprehension is present.

So, despite demonstrated defects in abstract and relational thinking, all the children in Healy's study had succeeded in mastering phonic generalizations and applying them to unknown words. Thus it may be concluded that hyperlexic children are not only able to read words through a visual matching-to-memory process, but are also successful at integrating visual and auditory, phoneme-grapheme correspondences. Hyperlexics, then, are "bottom-up" or totally text-driven readers, deficient in cognitive structures necessary for interaction between decoding and textual meaning;; they are uniquely responsive to external organizations while lacking ability to create organizational patterns of their own. Hyperlexics appear to be responding to syntactic constraints without meaning. Comprehension deficits appear to stem from a generalized cognitive disability in structuring incoming experiences.

The converse of this phenomenon would probably be that of 'perseveration', the name given to psychological behaviour which persists in using inappropriate schemata and fails to take into account disconfirming information (I am indebted to Dr J. Henzell-Thomas for bringing my attention to this point). In this case, top-down processing has taken over at the expense of any sort of bottom-up processing.

In both of the above cases - hyperlexia and perseveration - the reader is using a strategy which is appropriate for him, because it is the only one he is able to use in the circumstances. But insistence on the top-down or bottom-up nature of reading (whichever side one wishes to take in the argument) may result in inappropriate intervention.

2.3.2. Is comprehension unitary or manifold?

The emphasis in approaches based on factor analysis is not, as in psycholinguistic models, to come to an understanding of how we arrive at comprehension of text, but rather to see if, and what, subskills might be involved in reading.

2.3.2.1. Problems associated with factor analytic studies

As with the cognitive approach, so with the psychometric approach we should be aware of general criticisms that can be made (cf. Hewitt 1982; 13-14).

Firstly, the criticism that psychometric research has been *atheoretical* is perhaps the most damaging, since a comprehension test cannot be constructed without a theoretical model (cf. Goodman 1976). Psychometric research has been atheoretical in two senses: in that researchers have never fully articulated a theoretical rationale for

the supposed sub-skills which they set out to test, and in that the reading tasks used suggest a very limited theoretical view of what kinds of behaviour comprise reading. This lack of theory renders suspect the test's construct validity.

The reading tasks typically required by tests inadequately represent the wide range and level of possible reading tasks (such as reading a newspaper, following instructions, reading a brochure, a novel or a textbook etc.). This atheoretical approach thus leaves many questions unanswered and casts serious doubt on the ability of the tests to measure anything other than their authors' subjective impressions about reading.

A second important criticism of the psychometric approach is that the task of answering questions is itself untypical. Since it is quite possible that the processes required to answer questions are different to those required in typical reading and since questions can examine only a proportion of what could be comprehended in a text, it is likely that a research using the task of question answering will result in a very limited view of reading comprehension. It can never be certain whether it is the difficulty of the language in the passage or the questions which is being tested.

Thirdly, the psychometric principles of test construction demand that certain items, such as those everybody or nobody gets right, be discarded. But it is clearly possible that these items could be testing important aspects of reading comprehension while conversely those items retained may no longer be valid measures.

2.3.2.2. The subskill hypothesis

The most important recent study is that of Lunzer *et al.* (1979). They set out to answer the question: Is reading comprehension unitary or manifold?

Their operational definition of comprehension was used as the rationale underlying the use of comprehension tests as a measure of effective reading and which were meant to find out whether the several tasks which may be exemplified in comprehension-test items derive from distinct skills or subskills or whether comprehension is a unitary ability (p.39). They point out that the persistence of the concept of separate subskills derives partly from the weight attached to the earlier work of Davis (1944) and partly to the attractiveness of *the subskill hypothesis*, i.e. the notion that reading necessarily consists of a variety of component skills.

The strength of such an analysis would be, according to Lunzer *et al.*, that if the results had turned out to be positive, the profile which emerged would constitute a

diagnostic tool with the following elements:

- words in context;
- literal comprehension (the equivalent of 'direct reference');
- drawing inferences from single strings;
- drawing inferences from multiple strings;
- interpretation of metaphor;
- finding salients or main ideas;
- forming judgements (a kind of 'evaluation', except that the reader is not required to make value judgements about the worthwhileness of the passage).

This categorisation was conceived of as "partly hierarchical and partly as corresponding to very clear differentiations" (Lunzer *et al.* 1979;45)

No evidence was found for the existence of these separate 'skills'. Moreover in an analysis of the *Edinburgh Reading Test* (stage 3), which consists of five separate subtests, each using a consistent question mode, each bearing on two or three different short passages, and purporting to measure, respectively, facts, sequences, main ideas, viewpoints and vocabulary, it was found that one single factor accounted for 81% of the total variance.

Thus the hypothesis that the several tasks used in tests of reading comprehension call on distinct subskills which can be differentially assessed and taught had to be rejected. (Lunzer *et al.* 1979;59). The results of the whole experiment, then, "... would seem to be entirely consistent with hypothesis of a unitary aptitude of comprehension and do not conform with the hypothesis of two levels of comprehension skills <higher and lower>" (ib. p.62).

Lunzer *et al.* do not, however, dispense with the notion of 'skills' entirely, but rather place it in a different perspective: skills, for them, are different comprehension tasks which describe the sort of questions that one can and should include in a varied and interesting comprehension test which is "an indirect measure of the adequacy of reading" (p.68) The negative finding about skills is quite consistent with the view that comprehension tests may serve a useful purpose and also that the best tests of comprehension will include a variety of tasks: "A good comprehension test is a (necessarily indirect) measure of a pupil's ability to reflect on what he is reading. It is also a stimulus for such reflection." (p.69)

It is this "willingness to reflect" which is at the heart of the question for Lunzer *et al.* ; reading comprehension cannot be broken down into a number of distinct subskills. The evidence pointed strongly to a single aptitude and no support was found for the hypothesis that some pupils might "possess" lower-order skills but not higher-order skills: "... individual differences in reading comprehension should be thought of as differences in the willingness and ability to reflect on what is being read. This is, of course, not a simple characteristic, nor is it innate. it is the outcome of many factors, including reading fluency, intelligence and interest." (p.300)

2.3.2.3. Reading skills or cognitive skills?

Harrison and Dolan (1979), in a smaller scale study, pose similar questions to Lunzer *et al.* and ask: "...is it feasible to consider reading comprehension in the context of 'reading' skills as opposed to one of cognitive skills which follow intitial decoding?" (p.14) This, it will be recalled, was one of the issues raised by a consideration of Sticht's analysis.

Harrison and Dolan reanalysed Davis's (1944) data and also the *Edinburgh Reading Tests* (five subtests each consisting, as noted above, of a putative subskill) and discovered the emergence of a single factor, with hints of other factors in the case of the *Edinburgh* tests. But since each subtest had very different types of testing procedure and item structure this was thought to relate more to item-specific differences than to subtest content. Harrison and Dolan conclude that classifications and taxonomies of comprehension skills must be treated with caution, since methods have yet to be devised to measure them reliably.

2.3.2.4. Classifications of subskills

Rosenshine (1980) has a review of some of the earlier factor-analytic studies and also examines a range of sources in the search for some consensus as to what might be a reasonable classification of putative subskills. He recognises three main types: The first, 'locating details', is the simplest and involves recognition, paraphrase and/or matching.

The second group might best be labelled 'simple inferential skills' and refers to the ability to draw inferences after reading short segments of a passage (including, for example, understanding words in context, recognising the sequence of events, recognising cause and effect relationships, comparison and contrasting).

The third group might be labelled 'complex inferential skills', as , for example,

recognising the main idea/title/topic, drawing conclusions, predicting outcomes. But, as Rosenshine points out, there is often considerable difficulty in distinguishing among these inferential skills; this is recognised in the taxonomy developed by Barrett (in Clymer 1968) for example, where both complex and simple inferential skills are placed on the same level, namely that of inferential comprehension. Rosenshine also found a number of 'unique skills' (i.e. skills identified by some authorities but not by others) such as 'distinguishing between fact and opinion' or 'determining author's purpose'; these are more than simply different words for what others recognise – it would be difficult, for example, to decide if these two skills represent simple or complex inferential comprehension.

The difficulties of deciding what a skill might be are enormous: "Each skill seems real and sensible. One can argue that some of the skills can be combined, but even then the list of unique skills would be over 30 ... One can also argue that some of these skills should be split or arranged according to subskills." (Rosenshine 1980;539)

As a result of his analysis, Rosenshine concludes that across several sources there is a consensus that reading comprehension entails about 7 skills:

- recognising sequence
- recognising words in context
- identifying the main idea
- decoding detail
- drawing inferences
- recognising cause and effect
- comparing and contrasting

However, if one also includes the unique subskills then the total number of possible reading skills is in the hundreds. (Rosenshine 1980;540)

Analyses of Davis's original 1944 data reveal similar problems. Harrison and Dolan (1979) remind us that on exactly the same set of data, Davis found five factors, while Thurstone (1946) found just one. In 1968 Davis, using a 'uniqueness analysis',identified five unique skills, though he appears to have been looking for eight: Recalling word meanings, finding answers to questions asked explicitly or in paraphrase, drawing inferences from the context, recognising purpose, following the structure of a passage.

But Davis (1972) using factor analysis found four skills : Recalling word meaning, determining meaning from context, finding answers and weaving these ideas together in context, and drawing inferences from the content.

Spearritt (1972) used 'maximum likelihood' factor analysis and unearthed four factors/skills: recalling word meaning, drawing inferences from content, recognising purpose, and following structure. Of all the skills, vocabulary ("recalling word meaning") was best differentiated. Spearritt concluded that although certain comprehension skills can be differentiated present types of reading comprehension tests, as distinct from word knowledge tests, largely measure *one basic ability*, which may well correspond to the label of "reasoning in reading" (Spearritt 1972;110)

Finally, Thorndike's (1973) analysis,using reliability coefficients, concluded that the reading skills selected by Davis were not distinguishable and also claimed that the distinction between 'word knowledge' and 'reasoning in reading' (or inferring from the text) was not justified because there was little differentiation between word knowledge and paragraph comprehension in the factor analysis (Rosenshine 1980;543).

A study by Zuck and Zuck (1984) has investigated how we might identify the skill called "recognising the main idea" in EFL. A specialist text on a currently controversial issue in biology was taken , and specialist and non-specialist native and non-native speakers (i.e. four groups in all) were asked to provide questions relating to the main ideas of the passage. Several differences between the groups emerged; the answers to questions posed by non-specialists tended to be more localised in the text, often based on the information in a single sentence or single paragraph.

In contrast, the specialists tended to ask questions based on an interpretation of larger units of the text, or even the text as a whole; the specialist questions required inferences more often than the non-specialist questions. The non-specialists appeared to use the "rhetorical significance" as one means of simplification.

The ESL teachers tended to use questions such as : "What is the author's purpose in including the first two paragraphs?" in order, it was reported, to help their students learn what to ignore on the detail level.

In addition, the native-speaker/non-specialists reported that in the selection of key words they relied on the text itself, while the other groups utilised various reader models: the NS/specialists tended to list words which a 'lay' reader would not find in a dictionary. The non-NS/non-specialists tended to list words which they had to look up themselves.

In attribution of difficulty, specialists cited the cause of difficulty as complex specialised concepts, and the non-specialist ESL teachers cited complex linguistic features.

Zuck and Zuck (1984;135) conclude that the data from the sample suggest that there may be systematic variation between specialists and non-specialists in requests for definitions, recognition of certainty of claim, local versus global questions, implicit versus explicit information, and explicit use of rhetorical information to simplify.

In fact, this does not seem all that surprising, but if ESL teachers turn out to focus on linguistic difficulty and specialists tend to focus on subject difficulty in devising test questions for the same passage, then ESP 'communicative' testers should be wary of claiming that they test more than simply linguistic knowledge, even if it be through the medium of specialist texts.

In direct and deliberate contrast to Rosenshine (1980), Hillocks and Ludlow (1984) claim not only that reading skills can be isolated but that they are also hierarchical. Their model is based on two assumptions: first, that answers to questions represent skill types and, second, that a question must be classified as a skill type in conjunction with the text from which it is derived. The model is divided into two major levels: literal questions (those whose answers appear directly in the text) and inferential questions (those whose answers are cued in the text but are not stated therein) - "It is simple logic that if readers cannot retrieve information that is stated directly in the text, they will not be able to make inferences from that information. Thus the two major levels are taxonomic, at least logically. In a similar fashion each skill type can be discriminated from the others in the set." (Hillocks and Ludlow 1984;8)

At the literal level of comprehension, question types/skills relate to:

- Basic stated information
- Key detail
- Stated relationship

At the inferential level of comprehension, the skills relate to:

- Simple implied relationship
- Complex implied relationship

- Author's generalisation
- Structural generalisation

(pp.9-13)

The study analysed scores from tests on the Rasch rating scale model and concluded that items were hierarchically and taxonomically related to each other; readers who are incapable of answering lower level questions will be incapable of answering higher level ones, while those who are capable of answering higher level questions are also capable of answering lower level ones. "Does lower level comprehension enable upper level comprehension or *vice versa*? ... The most plausible explanation ... appears to be the former." (p.22)

2.3.3. The problem of hierarchies in problem-solving and in identifying reading skills

A concern with reasoning in reading and with the notion of skills leads directly to the problem of hierarchies: Can hierarchies of skills or reasoning processes be established, and how?

In an analysis of many examples of learning hierarchies so far published, Horne (1984;164) has shown that not only are hierarchies "conceptually and methodologically unsound", but also there is no source of exact domain order and there is no methodology for validating domain order: "learning hierarchies in the present state of the art must not be used as a basis for diagnostic tests nor as the source of test construction theory." He goes on to suggest that without a detailed knowledge of the nature of the skill continuum no measure of construct validity is possible.

Educationally, the notion of a (causal) hierarchy of learning skills in *any* field has been severely criticised by Horne (1983); certainly, the idea of a *discourse* hierarchy is both over-comlicated (Brown and Yule 1983) and untenable *a priori* (Jackson 1984). Nevertheless the idea is a seductive one; Blanton (1984) describes a model to teach academic reading to advanced ESL students which operates on the "simple premise" that written English discourse is "conceptually hierarchical" (Blanton 1984;37)

This makes little sense; what it seems to mean is that there are 'notional units' within a text which are related by subordination and coordination, understood as "integral structural features of the text." We are led by a devious and verbose route to the conclusion that "the expository texts the students were reading did not consist of words, sentences and paragraphs, but ...blocs of informational language" (op.cit.;40). It is a strange text that does not consist of words, sentences and paragraphs.

It is true, as Blanton points out, that the linear visual arrangement of written language should not be mistaken for an actual linearity of thought, but this does not require a commitment to some tenuous hierarchical abstraction. If 'hierarchy' means no more than 'organisation' then there is little objection. but if it takes on the conception of a discourse tree structure then it is a much more controversial idea (cf. Morgan and Sellner 1980 for example).

To approach reading from reasoning is to encounter the problem of tasks. Blanton (1984) suggests that 'prediction' is a "process of selecting and rejecting hypotheses, similar in many ways to problem-solving" (p.41). The trouble with this, and any approach to reading which emphasises 'skills', is that it elevates a particular test task to the status of a construct.

Almost all discussions of hierarchies in this context go back to the work of Gagne (especially Gagne 1962). Gagne's definition of 'knowledge' is: "that inferred capability which makes possible the successful performance of a class of tasks that could not be performed before the learning was undertaken." (1962;355)

To establish a hierarchy of tasks the question is asked of the final task: "What kind of capability would an individual have to possess if he were able to perform this task successfully, were we to give him only instructions?" Repeating the procedure defines a hierarchy of subordinate knowledges growing increasingly 'simple' and at the same time increasingly general.

The fundamental flaw in applying this idea to text and understanding of text is that while mathematics consists of and is defined by tasks i.e. one cannot but proceed by the solving of problems presented, text is not in itself a task; tasks can be created from text, but different tasks will put a different form or shape upon the text. Quite clearly "Solve $x^2 - 2x - 8 = 0$" is a task in a way that "Read/Understand: He larved ond he larved on he merd such a nauses The Gracehoper feared he would mixplace his fauces" is not, though one could construct tasks from it.

A concern with reasoning in reading should not then suggest that certain tasks reside inherently in the text.

2.3.4. Operational definitions of comprehension

Under this heading we include any approach to or definition of reading that is not based on or not concerned with empirical evidence or support. Thus, an example would be Bormuth (1969): "Comprehension is thought to be a set of generalised

knowledge-acquisition skills which permit people to acquire and exhibit information gained as a consequence of reading printed information."

The following broad categories can be discerned:

2.3.4.1. Comprehension as a response to the language system.

Specifically, comprehension is not just a set of mental processes which can be defined independently of language. Rather it is a set of processes which operate upon specific features of the language. (Bormuth 1969;50). Notice that here we are concerned with something called 'comprehension' rather than 'reading' as such, and at such high levels it is more or less essential to have an operational definition if we want to make progress, since an explanation of what it might mean to 'understand' something could lead into complex epistemological problems.

2.3.4.2. Comprehension as a statement of success.

Lunzer *et al.* (1979), although they approach the question from a radically different angle, are also concerned with operational definitions of comprehension, though they contrast sharply with Bormuth when they discuss the notion of reading comprehension in the light of Ryle's distinction between a 'got it' word and a 'doing' word: "Comprehension is not a label or a description of anything that the reader actually does; it is a statement about how well the thing is done. And it relates to the whole." (p.67). This analysis of the concept "comprehension" incidentally highlights the difficulty inherent in answering Davies and Widdowson's question posed at the beginning of this chapter: "What is is that we do when we fully comprehend written communication?" We do not "do" anything; the question therefore is in these terms strictly unanswerable (cf. Ryle 1954; 102).

Lunzer *et al.* also point out that there are at least two levels on which understanding ('comprehension') may operate: At the lower level it is sufficient that the reader satisfieshimself that that the matter which he reads makes some sort of sense. To do this he must know the meaning of most of the words and he must see that they hang together grammatically and conceptually (pp.37-8).

But beyond this level, and to enable the student to learn by reading, the reader must penetrate beyond the verbal forms of the text to the underlying ideas; so we arrive at a definition of reading calling for the reader "... to penetrate beyond the verbal forms of the text to the underlying ideas, to compare these with what one already knows and also with one another, to pick out what is essential and new, to

revise one's previous conceptions." (p.38) Of course, the problem with this is that it is unobservable and we are therefore forced to rely on operational criteria. "... answers [from comprehension questions] ... will then constitute an operational criterion of [the pupil's] comprehension ...What we have done is to translate a psychological or pedagogical goal ... into a set of behavioural objectives." (p.39)

2.3.4.3. Comprehension as 'something beyond word recognition'.

Catterson (1979) in arguing for a discourse analysis model of reading also relies on a definition of some generality, though in this case we seem to be approaching desperation: "... comprehension in reading depends on ... a word recognition factor and ... *something beyond word recognition* ... Although we cannot explain exactly what reading comprehension is, over the years we have gained insights into the kinds of practice that seem to produce the desired results."(p.2)

2.3.4.4. General criticisms of reading research

The two research paradigms - the cognitive and the psychometric - do have central features in common, and these features are open to criticisms which concern some of the fundamental issues arising in the behavioural sciences (cf. Hewitt 1982; 15-18).

Firstly, there is the problem of *measurement*. In order to measure something one must have a fairly clear idea of what the phenomenon is and how it manifests itself. In other words, until a theory about reading comprehension has been fully articulated and preferable some empirical support has been produced for the theory then the validity of the measure of comprehension must be in doubt (cf. Goodman 1976). Since no such complete theory of comprehension exists, measures such as question answering and recall are at best indicants.

Associated with the problem of measurement is the use of significance testing to support inferences. Carver (1978) points out that it is practically impossible to determine whether a difference obtained on an educational measurement device is significant in the sense of being important. In addition, significance testing can only be used if random procedures have been followed in choosing samples or assigning subjects to treatment groups. If random sampling did not occur, then statistically 'significant' differences could have been due to many other factors and therefore the test of significance is irrelevant. Again, inferences based on statistical tests of significance cannot be made beyond the specified samples from which the sample was drawn and cannot be made beyond the particular contexts of the experiment.

The second major criticism concerns the *theory* of reading comprehension adopted by the researcher. In order to construct a measuring instrument a theory of the competence to be measured is essential. One reason for inadequate hypotheses is the paucity of deductive theory which explains, predicts and specifies the necessary and sufficient conditions to any given phenomenon.

Thirdly, the *generalizability* of reading comprehension research results is limited by the following: the lack of theory and hypotheses as a solid foundation for empirical research; the failure to describe adequately the characteristics of the sample and the population from which they have been drawn; the reliance on what are often inappropriately applied statistical tests of significance as a basis, actual or implied, for scientific inference; and the lack of replication results.

Given the soundness of some of these criticisms (and others discussed earlier) what can be done? It has been suggested (Hewitt 1982; Filstead 1970) that research on reading comprehension needs to move towards a more 'qualitative' orientation in data collection and analysis. This would involve a reconceptualization of the notion of reading comprehension, conceiving it as more than just a cognitive process if it is to be understood in its full complexity. We would need to take account of the social , cultural and political contexts in which it occurs and how these influences affect how readers approach texts, their attitudes to reading, their comprhension Such indeed is the direction we would follow if we began to adopt more of the methods that have been used in literary criticism for many years now (cf. Rosenblatt 1978 e.g.). The discourse analysis tradition however shares many of these insights, and it is to this type of analysis that we now turn.

2.3.5. Frameworks for comprehending discourse

2.3.5.1. The problem of meaning

The literary criticism tradition to which we have just referred centres around arguments of where meaning is to be found: in the text, in the reader, in the author, or in some interaction between any or all of these. Such aspects of meaning in reading are well rehearsed in the schools of structuralist criticism, where the plurality of significations in a text and the equal validity of each have been investigated. Barthes, for example, says that "to read is to find meanings"; and Sartre has it that "reading is directed creation". The central question is whether there is simply one (usually authorial) meaning or a diversity of textual meanings to be found and explained; the text is not viewed as a neutral object: "The text in its words contains

the code that allows the reader to seize its meaning, but the text's meaning can materialize only in the reading of it" (Kronik, n.d.; 44).

This is not to say that we have to venture into the realms of literary critcism to find similar insights within applied linguistics. Indeed, the whole tradition stemming from Bartlett (1932) represents a 'scientific' attempt to investigate the same problem of meaning. The most important and influential study is that of Anderson *et al.* (1977), where the notion of 'schemata' is analysed and investigated.

The 'schemata' view of reading goes beyond purely linguistic matters, and in a sense is opposed to the view that deficits in either language or in reading skill are responsible for deficits in reading comprehension. The argument is that it is simply assumed that knowledge can be expressed in printed language and that a skilled reader can acquire knowledge from reading. On this view, each word, each well-formed sentence, and every satisfactory text passage 'has' a meaning. The meaning is conceived to be 'in' the language, to have a status independent of the speaker and hearer, or author and reader. A failure to comprehend a non-defective communication can in principle always be traced to a language-specific deficit. This is a theorem that follows directly from the axioms that the skilled reader can decode the language into knowledge. Therefore, it is assumed, difficulties in comprehension can be traced to failures of skill. Some of the words may not be in the reader's vocabulary; a rule of grammar may have been misapplied; an anaphoric reference may have been improperly coordinated, and so on.

However, Anderson *et al.* (1977) were able to show that the meaning of a communication depends in a fundamental way on a person's knowledge of the world and his/her analysis of the context as well as the characteristics of the message. In other words, comprehension of words, sentences and discourse can not be simply a matter of applying linguistic knowledge; every act of comprehension involves one's knowledge of the world as well - the knowledge structures which the reader brings to the text.

This kind of conclusion has been found again and again whenever studies have been done on the knowledge, beliefs and values in readers which influence comprehension processes (cf. e.g. Reynolds *et al.* 1982). However, demonstrations of the role of individual differences in comprehending a text based on differences in the availability of schemata pose a general problem: comprehension could not be so adaptive if it were too strongly determined by schema-based expectancies; how, for example, could speakers or writers effectively convey information if receiver

interpretation were such an idiosyncratic phenomenon? In fact, writers do have a variety of techniques for influencing readers' interpretation of discourse; we could say that one way such techniques can have their effect is by eliciting the appropriate interpretive schema in the reader's mind. Pratt *et al.* (1981) have clearly shown that ambiguous passages such as those used by Anderson *et al.* (1977) can be manipulated by the judicious use of italics to force one interpretation or another. It seems, in fact, that such passages are not ambiguous unless we go out of our way to make them so. Such, of course, is the case with sentences of the type 'Flying planes can be dangerous'.

Moreover, it has to be said that the texts upon which such studies are based tend to be extremely contrived – as if deep insights into the nature of human comprehension could be gained from subjects' responses to such chestnuts as "Visiting aunts can be a nuisance". The crux of the matter is our notion of 'meaning'.

2.3.5.2. Universals in reading

The argument from the discourse analysis side will be that meaning does not reside in sentences in isolation but only from the 'value' they take on beyond their 'signification' in continuous text. While this is true, there is an ever– present danger that analysis that is appropriate to the processes of general cognition may be transferred inappropriately to specifically linguistic constructs.

This question relates to the issue of identifying what is universal in the reading process: are skimming, scanning, predicting etc. specifically language–bound (and as such therefore in need of being taught and trained in the foreign language) or are they merely manifestations of certain cognitive activities which would operate regardless of the particular language or even language mode (e.g reading or listening) in which they are required?

Morgan and Sellner (1980;175) put it thus: "The main issue is quite simple ... How much of the competence that underlies the ability to understand and construct discourse is specifically linguistic, and how much is just the manifestation ,in use of language, of mental systems more general than linguistic competence?"

In discussing the question of speech acts, Morgan and Sellner make reference to the inferential system that gives rise to conversational implicature or indirect speech acts and remark that "a good amount of meaning (in the loose sense) is conveyed, according to this picture of things, by means of inferences about the speaker's intentions and purposes. But if Grice is right, these inferential principles are just the

application, to use of language, of principles that are quite general, and found in other areas of human interaction. Insofar as this is true, the mental systems underlying indirect speech acts are not linguistic systems." (p.176)

They go on to examine the notion that linguistic theory can be extended to account for discourse problems, a superficially attractive idea because, for one thing, discourse seems to exhibit properties that are similar to well-known properties of sentences (topic-markers, pronouns, beginning/ ending markers etc); but the similarity between text properties and grammatical properties is an artefact of superficial inspection, and Morgan and Sellner are forced to the conclusion (p.181) that "with at least some kinds of text structure, the ability to impose some kind of organizational structure ... is most likely the same ability one uses in imposing structure on observed reality. In fact, it is tempting to suppose that what is called 'pragmatics' is just the application, to verbal problems, of very general abilities for interpreting the everyday world ..."

2.4. The dimensions of reading in L2

If it is difficult enough trying to identify the dimensions of reading in L1, the problem is compounded in L2 reading by the addition of other factors.

Part of the problem with L2 reading has been a refusal to face up to the questions involved; Oller (1979; 232) abrogates all responsibility: "... perhaps it is best to call the item type a "sentence paraphrase recognition" task. Thus by naming the task rather than positing some abstract construct we avoid *a priori* validity questions." Or again: "Another way of referring to it is to say that it is a task that requires reading and answering questions - leaving open the question of what the test is a test of." (Oller 1979; 235)

Others have tried to be a little more analytical; Heaton (1975; 103) distinguishes ten different levels or aspects of the reading skill, ranging from simple letter recognition through gradually more complex cognitive elements to the advanced stage of critical reading with the reader employing a variety of strategies. Valette (1977; 166) on the other hand seems to recognize only two basic elements: structure and vocabulary. Davies and Widdowson (1974; 155), in addition to the motor skills, emphasize the cognitive nature of reading, the reader "obtaining meaning from print"; they resolve the "single coordinated activity" of reading into two broad areas: knowledge of the language system, and a knowledge of the way the system is used for the purposes of communication.

Clearly, there is a need for clarification.

2.4.1. The adoption of L1 models

Models or assumptions from certain areas of L1 reading research are often taken without further thought. Eskey (1973) for example appears to take over the views of Goodman and Smith that the fluent reader is "a person who can make optimal use of all the redundancy in a piece of text", without considering the possibility that this may simply not apply in the case of L2 reading. Brumfit (1978) also appears to place his hope in psycholinguistic research: "A notion of reading as a process of creative interaction with the text, with the reader predicting the message increasingly accurately ... provides the beginning of a theoretical basis." Though, as Brumfit rightly points out, we rely on a number of areas of study which are not directly involved with the teaching situation at all, so that 'reading' occupies an undefined position midway between linguistic and cognitive studies. Brumfit also feels that it is unclear whether advanced reading work in a foreign language poses particularly different problems from advanced reading work in the mother tongue.

Ulijn (1984a) has clearly swallowed the psycholinguistic bait whole: "We consider the reader as a system structured with different subsystems, and we are concerned with an ever-increasing processing capability." This sort of gobbledegook verges on the parodic; what it leads to is a dehumanized view of the reader: "... we interpret the linguistic process of reading comprehension as the performance of a text-analysis programme" (Ulijn 1984b;67). Nevertheless, Ulijn's conclusions do permit a relatively simple interpretation: the transfer of native language to foreign language reading in a positive or negative way is much more important on the lexical than on the syntactic level (or if you prefer: "these experiments support a conceptual strategy of the text and sentence parser in normal comprehension-oriented FL reading" !).

The problems with this kind of approach have been outlined by Meara (1984). He points out that a great deal of the current research on reading in a second language is concerned with higher-order processes such as discourse handling, reactions to textual cohesion and so on. The logic of the research is roughly:

1. Choose a model of the reading process, usually based on L1 reading, and decide in which particular part of it you are especially interested.

2. Find a robust, well-researched experimental paradigm which is generally taken as providing support for the model you have chosen and for the importance of the special subcomponent in the model.

3. Replicate an experiment which shows this effect, but use a group of non-native speakers as well as a group of native speakers as subjects.

4. If you are lucky, then the learner group will fail to show the expected effects; this allows you to infer that the learners are insensitive to the the particular features that your experiment manipulates, and to argue that any model of reading in L2 must take this deficiency into account.

Part of the problem here is that too many assumptions are made about lower order processes, not to mention the criticisms that can be made of the L1 'cognitive' approach.

2.4.2. Is there a difference between L1 and L2 reading?

2.4.2.1. The reading universals hypothesis

Goodman (1970) has proposed that the processes involved in reading will be much the same for all languages with minor variations to accomodate the specific characteristics of the orthography used and the grammatical structure of the language; the basis for this claim lies in the assertion that "semantic aspects of the reading process *cannot* vary to any extent from one language to another since the key question is how much background the reader brings to the specific reading" (Goodman 1970; 67). This would seem to be debatable. Kellerman (1981; 44-48) seems to assume this view too: "The ESL child is inevitably weak in ... the syntactic aspects of reading. His relative strength will be in the semantic region: his ability to infer meaning from context." There seem to be several unquestioned assumptions here.

A slightly different emphasis can be seen in Baltra (1983; 22): "From a cognitive standpoint it also makes sense to assume that reading for either entertainment or information are universal reading purposes. Thus, scanning, skimming, spotting central and secondary ideas, inferring non-explicit information, deducing meaning from context, comprehending the author's intention etc. would also seem to be reading activities employed by experienced readers in all languages." Baltra also sees rhetorical features as universals of some sort, based mainly on the claim that the processes and procedures of science (in this case) are the same no matter what the mother tongue of the scientist concerned, since scientific discourse represents a way of conceptualizing reality and a way of communicating which must, if it is to remain scientific, be independent of different languages and different cultures.

The advantage of viewing reading in these terms is that we escape the trap of assuming that because students cannot read in English they must know nothing about how writers convey information and how written information is most efficiently retrieved. As Swan (1985;11) so starkly puts it:"Exercises like this [role-plays designed to develop 'conversational strategies' which in fact teach discourse analysis] treat the learner as a sort of linguistically gifted idiot - somebody who knows enough language to express the (quite complex) ideas involved, but who somehow cannot put the ideas together without help. Normal students ... have the opposite problem."

Ulijn (1980; 22) has suggested that while foreign language reading teaching concentrates rightly on language factors, it may in this way risk underestimating possible reading universals common to L1 and L2 (such as oral, aural and silent stages in reading development). This problem has not been rigorously addressed in the L2 reading literature; the next section suggests how we might set about looking for an answer.

2.4.2.2. Two hypotheses

Alderson (1984) has the following two hypotheses which, he claims, represent the views one could hold on the relationship between reading in L1 and L2:

1. Hypothesis 1: poor reading in a foreign language is due to poor reading ability in the first language.

2. Hypothesis 2: poor reading in a foreign language is due to inadequate knowledge of the target language.

Two sub-hypotheses relating to the two major hypotheses can be made:

1. Hypothesis 3 (really 1a - a modification of H1 above): poor FL reading is due to incorrect strategies for reading that FL, stragegies which differ from the strategies for reading the native language.

2. Hypothesis 4 (a modification of H2 above): poor FL reading is due to reading strategies in the first language not being employed in the FL due to inadequate knowledge of the FL. Good first language readers will read well in the FL once they have passed a threshold of FL ability.

Little investigation of these questions has been carried out in a systematic way, though by posing the questions in this way it is suggested that answers will arise

from experimental psycholinguistic research, which will be open to the same criticisms of such research that we noted earlier.

2.4.3. Is L2 language proficiency unitary or manifold?

This is an important question for the implications it has for the status of reading. Lado (1961; 223) holds that reading in a foreign language consists of grasping meaning in that language through its written representation. The language difficulties that a student has in reading a foreign language are substantially the same as he has in understanding it aurally. So for Lado we can say in working terms that a student has learned to read a foreign language when he has mastered the specific difficulties of those of his language background and writing system. These difficulties will be related to interference from the native language and writing system. Reading is therefore one manifestation of the language, but not some sort of entity in its own right. This enables Lado (1961;228) to define reading in a foreign language as "the grasping of the full linguistic meaning of what is read in subjects within the common experience of the culture of which the language is the central part". Other types of reading are more properly the realm of reading in the native language (e.g. for literary appreciation). This seems to accord with some of the views expressed in the section above dealing with the reading universals hypothesis.

The arguments used to support the identification of subskills in reading are of the same type as those which are used to support the Unitary Competence Hypothesis (UCH) (Oller 1976; Porter 1983; Vollmer 1983). Both arguments rest on factor analysis or, which is much the same thing, correlational analysis, and may be sketched in outline as follows:

Reading skills: we name what we think are the subskills involved in reading and proceed to try to find these skills through the analysis of reading tests. Loadings on factors are presumed to be identifiable with the pre-defined reading skills.

UCH: we posit that the underlying variance in language tests can be accounted for by a single factor, which we shall name 'language competence'.

In a sense these arguments approach the analysis of data from opposite sides: the 'reading skills' argument posits the existence of identifiable skills, tests are constructed to look for those skills; the results, it is hoped, will show a number of different 'factors', which can then be said show to that independent subskills do indeed exist. The 'UCH' argument posits the existence of a single factor; we analyse test data based on the assumption that language skills can be decomposed into a

number of separate areas (reading, grammar, vocabulary, listening etc.) and hope that the results show the existence of just one factor.

If one accepts the *logic* of this style of argumentation (the results, in a sense, are irrelevant at this point) then it would seem likely that one could not hold as true at the same time the idea (a) that language competence is unitary and (b) that reading subskills can be identified.

In contrast to the notion of communicative proficiency as a complex entity is the idea of 'Unitary Competence' (cf. e.g. Porter 1983;190), or the 'Indivisibility' hypothesis. This states that there is a global language proficiency factor and no further special component proficiencies which can be separately tested.

The issue is complicated by the fact that it is not often clear in arguments for the UCH (e.g. Oller 1976) whether the argument is for global *language* proficiency or for some global *intelligence* factor. The factor *g* used by Oller to denote the global language proficiency factor is a deliberate echo of Spearman's *G*, which was supposed to account for individual differences in all tests of ability (cf. Nunnally 1978;507).

Stern (1983;352) points out that Oller's view represents a two-fold challenge to previously held views: the first is the idea of unitary language proficiency as opposed to the theory of proficiency as consisting of various components which combine differently in different individuals. The second is that expectancy is the key concept for such a unitary proficiency theory.

Criticisms of the UCH include those of Cummins (1979 and 1980) who suggests that theories of this type are based on test data but that language tests have a certain academic or cognitive character and that what in fact they test is a "cognitive/academic language proficiency" (CALP) and that they correlate highly each other and with intelligence tests because they have the same academic characteristics. What the tests fail to capture, it is said, is another quality of language use: basic interpersonal and communicative skills (BICS), which are essentially creative. BICS is by definition not tested, so we are left with CALP as the only testable aspect of proficiency.

Oller (1983) no longer holds a strong view of the indivisibility hypothesis. This is just as well in view of Pang's (1984) reanalysis of the Oller and Hinofotis data which revealed that:

1. The existence of a Spearman g-factor cannot be deduced.

2. The claim that a relatively high general factor suffices to explain the data is not warranted.

3. The residual matrix shows clearly that the second factor extracted is needed to explain the common variance in the data.

For Pang, the existence of a g-factor in language tests is still therefore an open question.

The real criticism of the UCH is the use it makes of factor analysis. The term 'g-factor' (i.e. Spearman g) refers to the one and only one common factor underlying all the variables in a set of data, so that, when the factor is analysed, all the variables would load significantly highly on this one g-factor. All other variation is therefore ascribed to specific factors unique to each variable. Vollmer (1983;14) points out that this type of analysis tends to overestimate the weight and significance of the first factor by not partitioning the total amount of test variance into common variance, test-specific and error variance; moreover a general factor can always be split into successively finer subfactors depending on the desired level of theorising, so the results obtained from factor-analytic investigations are definitely open to manipulation.

The extreme view of the UCH is represented by, for example, Flahive (1980), who examined the relationship between scores on a non-verbal IQ test, three reading tests (multiple choice, paraphrase and cloze) and TOEFL. His results suggested that traditional multiple choice reading tests are not simply tests of language proficiency but are also tests of non-verbal intelligence.

Apart from Vollmer's (1983) criticism of such studies, it should be noted that Oller and Perkins (1980;9) use the results of this and other studies to support "the fact that language skills of all the traditionally posited sorts are fundamentally related. This appears to be true for natives and non-natives alike." But surely no-one ever denied that the skills were related; it was always recognised that the sum of the whole was greater than any one of the parts (Davies 1978). Or as Spolsky put it: "we could find out about 'knowledge of a language' equally well when testing passive or active skills. This last does not mean that an individual's performance as a speaker is the same as his performance as a listener . . . All that it does claim is that the same linguistic competence, the same knowledge of the rules, underlies both kinds of performance."

The difficulty of distinguishing *g* from *G* is present in all discussions of language test data. Carroll (1972) has pointed out that comprehension tests may tend to be substantially correlated with intelligence tests, even those of a non-verbal character,

because reading and listening comprehension tests do not measure only what may be called 'pure' comprehension of language; they tend also to measure ability to make inferences and deductions from text content. The question is whether it is possible in fact to distinguish 'pure' comprehension of language texts from processes of inference, deduction and problem-solving that often accompany the reception of language.

Carroll (op.cit.;8) goes on to suggest that the problem we face is whether it is actually useful to draw a line between 'simple comprehension' and 'inferential processes', and if so, where on the continuum the line should be drawn.

The relevance of this to the case of reading comprehension tests is that testing comprehension is going to involve at least two conceptually separable stages of the comprehension process: we would like to find out, in a given case, the extent to which the individual 'correctly' apprehends the purely linguistic information that is 'committed' to the message, and also the extent to which he 'correctly' relates that information to some wider context. (Carroll op.cit.;14)

Proficiency tests, however, fragmented into whatever aspects of language it is deemed appropriate to measure, may be reliable measures of overall language ability. Hisama (1980;48) looked at batteries of tests in English as a second language and concluded that the results of several tests are likely to give a better prognosis of college achievement than any single test. The reason for this was said to be that while the rationale for multiple measurement is that any human ability is complex and thus requires a wide range of item samples and that therefore a single set of test items or a single test may fail to measure a very complex human ability such as language proficiency, yet in the case of tests of English as a second language there seems to be a large amount of overlap in the information given by the subtests. Such overlap is demonstrated by the rather high intercorrelations usually observed among them.

2.4.4. Types of L2 reader

It is quite possible that our views of the dimensions of foreign language reading will be affected by the particular situations we have in mind. The following two broad groups seem to be identifiable.

2.4.4.1. Foreign language learners

For this group of learners, the second language is either studied as a school subject in its own right, much as French, for example, is studied in British schools. There is no question of the language being used for some immediate practical purpose. Or the language *is* being studied for some immediate practical purpose, but that purpose will be quite specific – reading foreign scientific and technical texts for example. For Ulijn (1984a; 71) the reader of a foreign scientific or technical text is hampered by the content words required for a conceptual analysis rather than by syntactic function words, which are experienced as difficult only if syntactic analysis is necessary. Beginning FL readers are said to need more syntax than advanced ones.

Kellerman argues that the child learning a foreign language is inevitably weak in the syntactic aspects of reading (Kellerman 1981; 48)

The important thing here is that the discussion focusses on quite specific elements of the language system: syntax and vocabulary especially. The priority of one or the other of them is in a sense subsidiary, since the main concern is with the language itself rather than what might be termed its 'rhetorical features'. Indeed, Ulijn (1984a;72) goes so far as to assert that studies on the textual level suggest that such textual cues as paragraph organization, definition and classification are similar in different languages; textual universals are "overwhelming" in LST (language for science and technology).

It would seem probable that only by ignoring any claim for the separate 'identity' of reading in a foreign language can we account for the fact that it is possible to learn dead languages. In spite of Meijers' claim that the course of Latin for theology students developed in the Netherlands is based on models of the reading 'process', it is quite clear that what has been produced in this case is a purely language-based course that happens to have found a way out of the problem of concentrating on vast masses of syntactic exercises before texts can be read (Meijers 1980; cf. also in this connection Heron's work on reading dictionaries, which combine grammar in the lexicon).

2.4.4.2. Second language learners

For this group of learners on the other hand, the second language has a status more or less on a par with the first language: foreign immigrants trying to function in a new country for example. Here, the specificities of need are less easy to identify and probably do not correspond to published curricula. Nevertheless, if the learners

are adult, we can probably choose to treat the language much as we would for our first category of learners. Children, on the other hand, may have to be taught such things as 'rhetorical function' in both the native and the second language.

2.5. The language problem

2.5.1. Introduction

To what extent is it true that we are wasting our time looking into discourse-level units as possible dimensions of foreign language reading? In other words, should we accept without a fight Swan's (1985;9) claim that "... language learners already know, in general, how to negotiate meaning. They have been doing it all their lives. What they do not know is what words are used to do it a foreign language. They need lexical items, not skills." This is the problem we turn to in this section.

2.5.2. Vocabulary

Meara (1984; 97) points out that it is very difficult to isolate higher order processes, such as the way learners handle discourse structure or even sentence structure, because we know so little about what goes on at the much more basic level of word recognition. It is often assumed theoretically that it is possible to manipulate higher order variables without worrying too much about more basic processes. But learners have considerable difficulty with words in their second language, even when they are dealing with words they know quite well; length and frequency of words have much the same effect on non-native speakers as they do on native speakers(and of course at different proficiency levels -cf. Perkins and Brutten 1983), but the main difference is that the effects are severely exaggerated. Meara reports that whereas English speakers appear to rely very heavily on the beginnings and endings of words in order to identify them, native Spanish speakers do not seem to react in this way; thus if Spanish speakers process words quite differently from the way native English speakers do, then it is very unlikely that speakers of other languages will do so too, particularly if word structure in these languages is significantly different from English. It seems reasonable to suggest that the performance of certain non-native speakers of English in taxing experimental tasks could be quite unrelated to what goes on when a native English speaker performs the same task.

In fact the question of vocabulary is often glossed over. Two of the most influential and well-respected textbooks dealing with the question of reading in a foreign language appear to be somewhat embarrassed by the vocabulary problem.

Nuttall (1982;65), for example, sets up a system of tackling reading problems through 'text attack skills', but has to preface the bulk of her book with the comment: "To deal with most of the reading skills it is necessary to assume that the reader's vocabulary is adequate. Otherwise it would be pointless to write about reading skills at all..." She goes on to say that while even moderate L1 readers can recognise 50,000 words, graded EFL readers seldom go higher than 3,500.

Grellet (1981) on the other hand seems to ignore the vocabulary problem entirely, and , furthermore, accepts Munby's (1978) list of reading skills without question and Goodman's model of (L1) reading: "Reading is a constant process of guessing" (p.7). Interestingly, Grellet (p.7) calls for a clear distinction to be made between teaching and testing, so that "Testing will obviously involve more accuracy-type exercises whereas through teaching one should try to develop the skills listed ..." But why, if reading is thought to consist of such skills as she analyses, should it not be tested as it is taught? Or perhaps she is tacitly agreeing that reading in the FL is more a language problem than a reading problem; she does not choose to pursue the point.

Salager (1983;55) reminds us that when the proportion of unknown words rises above the 10% level, reading comprehension sinks to frustration level. And, although she was investigating syntactic components of the FL reading process, Berman's (1984) study shows that many of the problems that have in certain quarters been ascribed to breakdown in discourse processing can in fact be seen as, at least partially, a problem of vocabulary: "words like *since, while* and *then* may always be perceived in terms of time relations rather than of reason, concession and result, respectively; some students always take *just* to mean *only a moment ago* in all its occurrences, others preferring to interpret it as *exactly* in all cases." (p.144)

In another study, Sim and Bensoussan (1979), defining reading comprehension as "the ability of the student to understand both content words (nouns, verbs, adjectives and adverbs) and function words (prepositions, pronouns, conjunctions and auxiliary verbs)" conducted an experiment to investigate the role played by content and function words in FL reading comprehension.

The experiment was based on evidence that cloze procedures have shown function words to be more easily replaceable than content words, and the assumption was also made that "students who have not fully mastered the reading skill that is needed to decode or interpret function words have more difficulty in reading texts than students who have mastered this skill, and that this is no less important a lexical skill than content-word decoding or interpretation" (p.37)

The conclusion of this experiment was that "function words as well as content word questions should be included in tests of reading proficiency." (p.40). This helps to suggest that knowledge of individual word meanings is more important in FL reading than overall 'discourse strategies'.

2.5.3. Syntax

To be able to read in a foreign language one does not, of course, have to be able to put explicitly linguistic labels to the parts of a sentence nor does one need to know how to analyse sentences on an overtly syntactic level. However, the knowledge of the syntax of the language must be passively known at least; otherwise how would the reader know which parts of the sentence relate to each other? Since we nearly always work in English or use research work which derives from the analysis of English we tend to assume that syntax is a minor problem, because the syntax of English is so 'simple'; we therefore allow ourselves to be drawn into arguments about the rightness or otherwise of concentrating on 'grammar', when we should really be concerned with 'communicative value'.

It is when we look at a more highly inflected language that we see just how important syntax (and grammar) can be in the reading process: and there is no need to erect a huge psycholinguistic apparatus to see this. Take these two lines from Horace's 'Soracte' Ode (I, ix) for example:

> nunc et latentis proditor intimo
>
> gratus puellae risus ab angulo

There are at least two ways of coping with this, depending upon one's 'knowledge' of Latin syntax: either we hold the individual words in our head, make tentative conclusions as to their relationship as we go along and then see the whole fall into place on the very last word; or we have to go back over the sentence and 'construe' it – in other words, our 'grammar' may let us down and we have to go back and analyse the sentence quite overtly. Higham (cited in Leishman 1956; 84) has commented on these lines that all the governed or attributive words come first and drive our minds onwards, encouraging us to 'go ahead' ... if our memories are not retentive, or if our knowledge of concords is insecure, we go back again and construe.

The case of Latin shows the problem of syntax at its clearest (for the British reader at any rate), but the problem can also be seen quite clearly in English. Take the sentence from Book I of Paradise Lost (ll. 84 ff.) which begin "From what highth

fal'n". C.S. Lewis comments on these lines that although the sentence is fairly complicated if you read it without bothering about the syntax you receive in their most natural order all the required impressions, but the complex syntax has not been useless; it has preserved the *cantabile*, enabling the reader to feel, even within these few lines, the enormous onward pressure of the 'stream' upon which he is embarked.

To what extent, however, is syntax a separate 'dimension' of reading in a foreign language?

2.5.3.1. Syntax with vocabulary

Of course, words in isolation are of no use unless 'propositional content' (Berman 1984;140) can be extracted from the context in which they occur.

Cooper's (1984) tests with practised and unpractised EFL readers in Malaysia concluded that "...unpractised readers were so preoccupied with the unknown word and its immediate context that they were often blinded to the meaning potential of the whole context offered."(p.128); furthemore, both groups were insecure and inconsistent in their understanding of meaning carried by tense, aspect, modality and non-finite participial clauses. In addition, Cooper found that "... unpractised readers are very uncertain of the *meanings* of sentence connectors" (p.132)

Once again, then, there does not appear to be strong evidence for a 'discourse' problem as such. Cooper's conclusion is that "...unpractised readers are severely handicapped by poor vocabularies" (p.133) especially with a high proportion of words that are common across subject areas (the so-called 'sub-technical' words), such as *contrast, similarity, function, characterise, depend on* etc. and also with a number of sentence connectors such as *despite, nevertheless* and *consequently*.

Salager (1983) reaches a similar finding in the context of ESP, particularly in the scientific context: "a more serious impediment to fluent comprehension are ... sub-technical words and ... academic vocabulary i.e. those context- dependent words which are used across different scientific disciplines, but which tend to occur infrequently in general word-frequency counts." (p.54)

2.5.3.2. Syntax in its own right

Ulijn and Kempen, cited in Alderson and Urquhart (1984a;12), claim that under normal conditions "reading comprehension is little dependent on a syntactic analysis of the text's sentences. It follows that second language reading comprehension is possible without mastery of the contrasting parts of the second language's syntax.

Usually the reader's conceptual knowledge will compensate for the lack of knowledge about linguistic contrasts between L1 and L2."

Kellerman (1981) maintains that the ESL child is "inevitably weak in ... the syntactic aspects of reading. His relative strength will be in the semantic region: his ability to infer meaning from context..." (p.44)

Hatch *et al.* (1974) conducted experiments with EFL readers to see how far syntax is involved in FL reading; they set out to answer the question: "Does the non-native speaker of English, who has less familiarity with English syntax and thus greater dependence on function words to signal syntactic meaning, actually pay closer attention to function words in such tasks than native speakers of English do?" (p.277)

Using an 'acoustic scanning' experiment they found that there was some reason to believe that for subjects not proficient in the language, perception of letters in the text is related to the subjects' processing of the syntax of the reading materials. "We could claim that such Ss not only saw the letters in the function words but also relied on the function words in order to understand the syntax of the sentence." The problem with this experiment, however, as the authors make clear, is that it was difficult to tell if the subjects were really reading or whether they were treating reading as a simple visual discrimination task; "... if we look at the scores for the comprehension tests which followed each of the cross-out studies, we find that the beginning level students, even though they marked letters in function words, were not able to use what they read in answering comprehension questions afterward." (p.283)

What evidence there is, then, is entirely compatible with the claim that reading in a foreign language is in essence **a problem of language**. For Yorio, cited in Coady (1979;9) FL reading difficulty can be traced to lack of knowledge of the target language; in certain relatively clear circumstances, however, it may be appropriate to dwell on the 'reading' rather than on the language aspect of the problem - "... for many foreign students the problem is not only to learn to read English, but to develop a reading habit for the first time in their lives ... the foreign student frequently suffers from a mental block ... a conviction that he must correctly process every word if he is to understand anything at all." (Eskey 1979;73)

2.5.4. Discourse structure

2.5.4.1. Schemata

We have already discussed the notion of schemata in L1 (Section 2.3.5 above). This has been a seductive idea for L2 analysts. Hudson (1982), for example, holds that much of the research into the L1 effects of schemata and context is applicable to L2 reading, mainly because schemata theory can partially explain the L2 'short circuit' of good reading strategies by proficient L1 readers. Steffensen (1986) invokes the notion of schemata to account for the fact that readers from different backgrounds bring their own cultural knowledge, beliefs and assumptions to the interpretation of a text and that they are likely to fall down on their 'processing' of the cohesive elements of a text if they fail to recognize that a text is about an example of a known class of situations.

The danger seems to be that either we are going to have to include within reading comprehension absolutely everything about a reader's life, background, hopes and expectations, or that we are going to become involved in matters of general cognition that can only confuse the issue. A choice has to be made as to how widely we want to cast our net.

2.5.4.2. Reading and cumulative errors

One question raised in a consideration of discourse structure is: to what extent is reading in a foreign language a 'cumulative' process, built from linguistic items no higher than the sentence? Or does there exist an identifiably distinct discourse framework, identifiable, that is, in the structure of the language itself, which is somehow different from general cognitive abilities?

The question arises from a consideration of such examinations as the *Cambridge First Certificate* and *Proficiency* examinations in EFL. For the writers of these examinations there is little doubt: reading comprehension is divided into two parts; first there are 25 multiple choice sentence completion questions designed to test 'usage' rather than 'use' and 'certain aspects of linguistic competence'.

More specifically, semantic sets and collocations account for 8 items, use of grammatical rules and constraints for another 8, with the remaining 9 items testing knowledge of synonyms/antonyms, semantic precision, adverbial phrases and connectives, phrasal verbs and prefixes/affixes.

There then follow three reading passages designed to test something more general (gist and 'language in use'). It is clear that it is the language system itself that is being tested through the reading **mode**, rather than anything identifiable as

reading in EFL.

Given our earlier comments about the importance of reading as a manifestation of the language rather than as an independently existing skill in the foreign language this seems perfectly acceptable; when it is put in such stark outlines, however, one is forced to ask how far reading is merely the accumulation of meaning gathered from sentences, rather than from the passage as a whole.

2.5.4.3. Inferential production

Frederiksen (1972;243) offers some useful insights into this question with his notion of 'inferential production': more super-ordinate kinds of processing tend to increase as comprehension of the passage increases. Thus the strategies that may be operating early in the perception of discourse may be different from the ones operating later. This may occur because the semantic resources that are available for stragegies to operate on have changed over the course of the discourse. As one gets well into the discourse passage, the frequency with which inferences begin to be made should increase surface structure may become less and less important later in discourse because one should rely more on assessing the significance of content by matching this against one's internalised semantic model of the passage. This concept would intuitively satisfy many of the feelings we have about extended discourse (and especially certain psycholinguistic ideas that meaning accumulates during reading) without forcing us to accept a discourse 'entity' as such. The pre-eminent linguistic status of the sentence as the processing unit would thus be restored.

2.5.4.4. The question of cohesion

Perhaps the clearest manifestation of the 'discourse' view of things in studies of reading in the foreign language has been the effort expended in teaching and testing 'cohesion' or 'cohesive devices'. Apart from the fact that any difficulty with 'cohesion' may in fact, as noted above, be the result of a vocabulary problem, the notion of 'cohesion' itself, especially as used by FL reading specialists is often uncritically accepted as some sort of property of the text, whereas this is in fact a descriptive device (one way of analysing text) and not an aspect of 'psychological reality' or an independently existing linguistic entity.

To illustrate the difficulties involved here, consider Nuttall's (1980) statement that "When such words [*he, our, this, them, they etc*] are used, they are signals to the reader to seek a meaning for them elsewhere in the text." (p.90) But as Webber (1980;147) points out pronouns do not just 'stand for' nouns ; just being capable of

constructing possible antecedents and referents for anaphoric expressions presumes complex cognitive abilities on the part of any understander. In other words, reference is to the world and not to the text; all too often in discussions of 'coherence' and 'cohesion' we take certain aspects of linguistic form as cause rather than effect of coherence.

Morgan and Sellner conclude: "As far as we can see, there is no evidence for cohesion as a linguistic property, other than as an epiphenomenon of coherence of content." (p.181)

Williams (1983) devotes a long article to teaching the recognition of cohesive ties in reading the foreign language, but his definition of cohesive ties would seem to preclude any value in this, if we assume that learners have any cognitive ability at all: "Textual cohesion is a semantic concept. It is concerned with semantic relations within a text ... such that the reader's ability to interpret a particular textual element depends on his ability to interpret another element. The elements are tied: thus we talk of cohesive ties in text. and inter-element semantic cohesion is one of the major features that enables a fluent reader to distinguish text from a random string of discrete sentences." (p.35)

Even here, the problem of cohesion seems often, on Williams' own admission, to be one of knowledge of vocabulary; he quotes (favourably) a study by Cohen *et al.* (1979) which showed that learners were not picking up the conjunctive words signalling cohesion, not even the more basic ones like 'however' and 'thus'. The informant noted that she had never known the meaning of 'thus', and had simply thought it marked off sentences. (Williams 1983; 39)

This is surely the crux of the problem, and not that something called 'cohesion' exists in the text rather than in the normal human cognitive activity of making sense of the world. Moreover, Williams posits what he sees as the three subskills which the efficient reader uses in dealing with discourse markers, but be offers no evidence that these are anything other than intuitive categories.:

1. First, the efficient reader recognises that a certain item is, in fact, a discourse marker (otherwise, the reader interprets it as just another word.)

2. Next, he must identify the function of the discourse marker concerned i.e. what type of proposition it is signalling. This functional identification enables him to predict the nature of the following information.

3. To assist him in recognition and functional identification, the efficient reader is able to draw on his knowledge of families (and sub-families) of discourse markers i.e.he knows that *for this reason* and *consequently* belong to the same family, but *in other words* and *nonetheless* to different families.

(Williams 1984;47)

In this analysis it is not at all clear how any of these three supposed subskills differ in any way from simple vocabulary recognition, and to say that "the reader interprets it as just another word" seems almost entirely devoid of meaning beyond something totally trivial.

Other statements seem equally difficult to interpret usefully: "A major problem is that a discourse marker represents an abstract concept, so that it is difficult for the learner to form a mental image of the underlying proposition being expressed" (Williams 1984;47) - in what way is the abstractness of the discourse marker different from the abstractness of language in general (particularly so-called 'function' words)? And how would it ever be possible for the reader , even the efficient reader, to form a mental image of any abstract proposition?

Williams is but one representative of a school of FL thinkers who follow a particular conception of discourse based on Halliday and Hasan's analysis of cohesion; although it might be easy to teach, it seems to be mistaken and not very useful.

In support of the idea that we are faced with a language problem (basically at the sentence level) rather than a 'discourse' or a 'reading' problem, one may cite Mitchell (1982;181-2): "... there is some evidence that fast and slow readers may differ in some of the higher level comprehension skills that are not associated exclusively with reading. Jackson and McClelland (1979) ... found that the strongest predictor of reading speed was the student's score on a listening comprehension test ... It seems likely that the procedures used to construct and link propositions are implicated in some way ... part of the difference between accomplished and less accomplished (mature) readers might lie in the efficiency with which they are able to link propositions."

2.6. Conclusion

To conclude this chapter we return to the four questions we posed at the beginning. "What do we mean by reading?" In spite of much intensive investigation into this question, it is hard to avoid the conclusion that a large part of the answer depends on the definition of reading that the questioner is presupposing. Even within

particular lines of investigation (comprehension, word recognition, cognitive processing) our current state of knowledge does not allow us to be dogmatic on any aspect of the reading process. The best we can say is that the various aspects of reading all need to be taken into account, and that over-insistence on, say, a 'top-down' view is not only likely to give a misleading account of what is involved in reading but is also likely to result in inappropriate intervention. Reading is in some sense a function of perception and cognition, but it is difficult to be more precise.

"Is reading in a foreign language different from reading in the native language?" Again, there is very little hard evidence to help us here. We suggest that on the whole it would appear that those who claim some sort of independent status for the construct of reading in a foreign language have the onus of proof upon them, if only because many of the arguments for this view tend to rely on some idea of 'subskills', the existence of which it is difficult to show. As with our previous question, the answer partly depends on what is meant by reading; a perfectly reasonable case could be made for including conventional tests of grammar in any test of reading, though the impetus to testing provided by communicative trends has meant that tests of reading in a foreign language have tended to simulate tests of reading comprehension in a native language. This aspect will be discussed in the next chapter.

"Is reading in a foreign language different from other activities in the foreign language such as listening?" Again, the paucity of evidence prevents us from giving a fully satisfactory answer to this question. On the basis of factor analytic studies, whose shortcomings we have noted, the answer would have to be a tentative 'yes'. The relationship between 'skills' such as reading, writing, thinking etc. in the native language, however, is not clear, and it would be difficult to maintain too strong a position on the divisibility of 'skills' in the foreign language.

"Is it possible to identify separate elements of reading in a foreign language?" This has not been possible in a first language, so there is no reason to suppose that it will be possible in a foreign language. This question is further explored in Chapter 7.

On the basis of available evidence, then, it seems that EFL reading could indeed be viewed as 'unidimensional', though initially this will be as much a matter of definition as of substantive investigation of content.

CHAPTER 3

ANALYSING 'ABILITY' THROUGH READING TESTS

3.1. Introduction

How should we set about ensuring that we are testing the 'ability' to read in a foreign language? Part of the problem is that of "conceptualizing attainment" (McIntyre and Brown 1978), and to that problem we turn in this chapter. Firstly, however, we must think about the content validity of our tests of reading. Lado (1961; 343) has argued that other fields of testing such as human intelligence, personality traits etc. do not have a body of content as neat and well-analysed as language today, and as a result such testing has had to rely heavily on statistically derived criteria for the determination and identification of factors; in foreign language testing, however, there is no substitute for content analysis: "we should probably fall short of our capabilities today if we were to begin from statistical factor loadings rather than from specific linguistic content." (loc. cit.)

3.2. Strengths and weaknesses of current tests

3.2.1. Types of test

An unresolved, perhaps unresolvable, conflict exists in the theory of testing between those who favour statistical criteria above all else and those who insist on the primacy of test content. It is an argument that has long been around: Horn (1966) and Ebel (1966) brought the matter into the full glare of public light in their published debate; Horn argued that while predictors must have internal consistency, assessments should have representativeness of content. Ebel, on the other hand, argued that *all* tests must yield variance of scores, otherwise what useful purpose do they serve?

While Ebel's view has important implications for practice, it seems unreasonable to ignore a rigorous examination of test content when this is feasible. To this we now turn.

3.2.2. Problems with existing tests

3.2.2.1. Language

Corrick (1984) discusses four examples from the English language comprehension paper of the London O-level examination in 1983 and comes to the conclusion that these tests do not measure an ordinary level of comprehension. This is primarily because the tests demand "the ability to comprehend a particular sort of language". This language is said to be "odd" in that it is full of metaphor and "elegantly tortured syntax". The test extracts examined are criticised on the grounds that they are inappropriate as examples of the common language of ordinary people; rather they belong to a specific genre, notably a sort of "sub-literary dilettantism", amusing to read but entirely peripheral.

It is this irrelevant complexity and conscious stylistic elaboration that is a danger with all tests of language that aim to test, in L2, 'advanced' reading. Such delight in complexity for its own sake can be seen quite clearly in, for example, the Cambridge Proficiency Examination and in most University translation papers.

A further danger here is that the questions themselves, in this case multiple choice selected response types, tend to be significant only because the language of the extracts is overly contorted; a crossword puzzle mentality is often needed to solve the tasks set successfully.

Fillmore (1982) also suggests that the testing industry has created a new genre for the written English language: "a genre whose characteristics are determined by very unnatural requirements of lexical choice, grammatical structuring, and synonym alterations, these dictated ... by the intention to test knowledge of particular vocabulary items ..." (Fillmore op.cit.;251).

Fillmore (1982;253-7) also showed how texts can actually change in the testing situation. Using the notion of an 'ideal reader' to help analyse what is required of the testee on a given text, Fillmore suggests that real readers differ from ideal readers in two directions: with respect to any given point in the text they may be *underqualified*, in that they do not know what the text assumes they know at that point, or they may be *overqualified*, in that they already know what the text introduces.

Using this method of analysis with adult readers presented with a segment of text at a time, Fillmore found that, for example, passages that are humorous when read all at once are not humorous when given out piece by piece. It is not just that the passages do not seem funny; sometimes their humorous intent is not even discerned.

Furthermore, if a text takes a digression and then returns to the main theme, the

return to the main theme may not seem very striking; in natural fast reading, by contrast, the digression itself would hardly be noticed.

This suggests that 'testing reading' must of necessity be limited to fairly low-level abilities (including knowledge of the language system) and can never realistically hope to capture what is involved in 'real' reading, for the reason that the very act of testing, with its associated fragmentation of the text, whether imposed by the tester or the testee, changes the quality of the text. In testing, we are never able to get in touch with "that ideal reader suffering from an ideal insomnia".

In the usual case, Fillmore (op.cit.;268) suggests that the reader has to know something about the real world in order to build on that to construct an *envisionment* (or coherent image or understanding of the states of affairs that exist in the set of possible worlds compatible with the language of the text) of the current text. In the case of the texts which he looked at, however, and in particular a constructed piece about the phonograph, "what we have to know in order to understand the text exhausts what the text tells us". This, claims Fillmore, is a clear case of a bad test, and most assuredly a bad test item.

This has implications for the choice of text in ESP tests; Fillmore's type of 'bad' text would be perfectly appropriate if we were interested only in testing linguistic knowledge.

3.2.2.2. Misuse of tests

Goodman (1982;289)) suggests that there are only two basic uses of reading tests which are legitimate:

1. To measure the effectiveness with which any person uses reading to comprehend written language. Within this the two main concerns are:

    a. flexibility in comprehending a wide range of materials; and

    b. degree of proficiency as compared to other readers or as compared to some absolute scale of proficiency in comprehending written language.

2. To diagnose the strengths and weaknesses of readers as an aid to planning instruction which will help to make them more effective.

Testing for each purpose will vary depending on the theory of the reading process and of reading acquisition which the tester uses.

A weakness of current reading tests is a failure to articulate views of the reading process and learning to read as a basis for building the tests, subtests and test items. Traditionally the successful reader is treated as a possessor of bundles of skills rather than as a user of written language; "semilogical sequencing criteria and hierarchical arrangements" are imposed on these skills, which are then isolated for ease in testing, outside any context of language use which they may have. (Goodman 1982; Hewitt 1982).

The practice of testing, however, has been found to be short of theoretical background and has been subject to some misuse. A recent study by Steadman and Gipps (1984) found, for example, that at primary school level in Britain the major use of test scores was for record-keeping, both for the primary school itself and for passing on to secondary schools at transfer. At secondary level the main use of test scores with a child on transfer was to assign children to teaching groups, while record-keeping was again a major use of of test scores obtained within the secondary school. Nearly all secondary schools used tests in the remedial departments in order to diagnose individual difficulties and monitor progress.

Steadman and Gipps (op.cit.;121) point out that record-keeping is essentially a passive use of scores and that rather than modifying the curriculum, teaching methods or their own assessments of children, teachers tend to look at test scores, think about whether they tie in with their own judgements, accept them if they do, ponder a little if they don't, and put them in the record book largely for the benefit of someone else.

The symbolic role of testing was also found to be crucial: it seemed that it was the setting up a testing programme that satisfied, rather than rigorous use of results, at leas so far as the LEA's were concerned. While for teachers, standardised tests provide a ready-prepared resource when preparation time is short. For headteachers, it is the power of comparison made possible by standardised testing that is attractive, as well as the apparent objectivity and neutrality of standardised tests which then offer a basis for discussing pupils with class teachers and parents. "Providing 'objectivity' for professional assessments is also very important for individual teachers" (Steadman and Gipps op.cit.;123).

At LEA level the same aparent qualities of objectivity, neutrality and comparability gave standardised testing a powerful appeal as a means of managing the system;

standardised tests appear to be impartial and precise.

In the end, comment Steadman and Gipps, it may be the simple ways in which the scores can be communicated, often in terms of a single figure, that make their attraction so all-pervading. They suggest that the detailed reporting of results from most criterion-referenced tests may make such test results difficult to absorb.

This is precisely the problem that arises in criterion-referenced test development in tests of EFL; Criper (1981) and Bruton (1985) point out (a) that authorities may not *want* detailed profiles of achievement and (b) that in opposition to the original test construction philosophy the ELTS test of EFL had, reluctantly, to provide an 'overall band score' to , in some way, provide a single interpretable score.

A further example of the misuse of tests comes from Steadman and Gipps (1984;124). They examined the use of the *Schonell Graded Word Reading Test* (GWRT) in schools; although the GWRT was first produced in 1945 and is certainly not in tune with today's reading goals "which stress in particular reading for meaning", it was found that in 1981 it was still being used by 46% of schools. Interrogation of the teachers revealed that its popularity is due to the fact that it is quick and easy to administer, easy to score, and, surprisingly perhaps, familiar.

Steadman and Gipps use this evidence to highlight the pragmatism behind much testing; a test may not be testing exactly the skill under examination, but if it is close enough and correlates reasonably well with the teachers' own assessments, then it will do. Tests with a diagnostic element were considered to be relatively complicated and time consuming, another factor contributing to the "... powerful inertia of usage once a test has become well known and widely used." (op.cit.;124) The simplicity of the GWRT made it attractive to teachers and at the same time contributes to its inadequacy as a measure of 'real reading' competence.

3.2.2.3. Statistical fallacies

One kind of statistical fallacy is to produce a single 'band score' by summing across different 'bands' of a profile, as in the example of ELTS above.

If a test is to be used for diagnostic testing then its effectiveness will be defeated if it is concerned more with the *quantity* of errors than with the specific phenomena revealed by performance on the reading tasks involved (the *quality* of errors).

If a score combines scores on 'skills' subtests with those on comprehension, then, since skills are ostensibly the means by which comprehension ('the end product of

reading') is achieved, such a score is meaningless (Goodman 1982;292). More generally, any fragmented aspect of language which is tested in the same place as a more integrative aspect and *which is assumed to be a component part of that more global skill* cannot validly be a part of a larger *single* score formed by summing the parts. Such a component test can only serve to increase the *reliability* of the test, by increasing the test length, but does not add any information for diagnostic purposes, at least so long as we insist upon a single score.

A further point is that any test constructed on norm-referenced principles is subject to the 'psychometric snare' noted by Popham (1978): namely, that in time an achievement test is likely to become an aptitude test. Popham (1978;83-4) points out that because the purpose of a norm-referenced test is to spread examinees out, 50% discrimination is preferred on items; teachers are therefore more likely to emphasise important topics, with the result that certain items will be answered correctly more often and thus excised from the test on subsequent revision, leaving only those that measure less important things. In time, the test items that spread people out best tend to be the kinds of item that are impervious to instruction. Such items, says Popham, are based chiefly on native intellectual ability and measure better what students bring to an educational programme, not what they leave it with. This argument assumes that test items are measuring *knowledge* alone, and not the application of knowledge.

3.2.2.4. Design problems

Goodman points to certain design problems in reading tests: first, there is the problem of *convergence*. Because there must always be a 'right' answer, at least in multiple choice tests, convergent responses (and convergent thinking) are rewarded at the expense of divergence - convergent responses always match the preconceptions of the test-maker.

Bormuth (1970;6) raises the same question in a slightly different context, pointing out that we seem to be in the position of having to accept the assertion that a test measures whatever the test writers claim it measures without recourse to definitive independent evidence. In the final analysis a test item bears a certain label just because the test writer and his associates say that that is what it measures.

Secondly, in multiple-choice tests no allowance can be made for the fact that the testee cannot show his misconceptions, when in fact a misconception may often be better than no conception at all. He is limited to the choice before him.

Thirdly, there is a problem that test-wise subjects may be able to use the format of the test to help them; a simple example would be that failure to answer one of the four questions required in the traditional British essay-type examination results in the immediate loss of a potential 25% of the marks available. Test-wise subjects know this, and answer four questions at all costs, even if the quality of answer is inferior.

Goodman offers no solution to these problems other than the asking of certain key questions to which he provides no answers:

- Can essential skills or strategies be isolated for testing without changing their relative values, their basic uses, or the reading tasks in which they occur?
- Are such strategies or skills universal across people, contexts, purposes, languages and orthographies?
- Is there an essential sequence in learning to read?
- How are reading skills or strategies to be understood in terms of how language works and is used?

Goodman's only comment on these issues as far as testing is concerned is that the diagnostic test of the future will be designed so that the strengths and weaknesses of learners will be made clear; in other words, he sees our salvation in the use of criterion-referenced tests.

3.3. Approaches to L2 testing

3.3.1. Traditional testing

Davies (1982;151) suggests that what remains a convincing argument in favour of linguistic competence tests (both discrete point and integrative) is that *grammar* is at the core of language learning: "grammar is far more powerful in terms of generalisability than any other language feature. Therefore grammar may still be the most salient feature to teach and to test."

3.3.2. Communicative testing

There can probably never be a 'strong' view of the 'communicative' position as far as the tester is concerned; this is because of the nature of the testing activity itself, which, as Weir (1983;93) says, is by necessity artificial and idealised.

The question of how to test communicatively often reduces to one of authenticity of materials, the argument against which has been fully expounded elsewhere (cf. e.g. Widdowson 1983). If the communicative argument is that we should avoid convoluted texts of the kind discussed by Corrick (1984), then there would seem to be little controversy in the matter.

Morrow (1981) points out that designing a communicative test involves answering these questions:

1. What are the performance operations we wish to test? These are arrived at by considering what sort of things people actually use language for in the area in which we are interested.

2. At what level of proficiency will we expect the candidate to perform these operations?

3. What are the enabling skills involved in performing these operations? Do we wish to test control of these separately?

4. What sort of content areas are we going to specify? This will affect both the types of operation and the types of 'text' that are appropriate.

5. What sort of format will we adopt for the questions we set? It must be one which allows for both reliability and face validity as a test of language use.

The problem with this approach is, paradoxically, that it tends to fragment language and communication, so that one ends up with a lot of little parts that have to be put together again - taxonomies based on Munby (1976) seem particularly liable to this fragmentation - whereas a principal tenet of language as communication must be that language is a unity.

Porter (1983) suggests that the testing of communicative proficiency is seen typically as a complex entity composed of a variety of high-level language abilities each needing to be tested separately.

Morrow (1983;117) is ultimately reduced to reliance on intuition: "My feeling at this stage is that we may have to face up to the fact that performance ('communicative') tests must remain an act of faith, but that their great virtue will reside in that extremely unscientific concept, 'face' validity."

In answer to the argument that communicative tests lack generalisability but grammatical tests do not, Morrow (1983;117) claims that 'linguistic' tests, paradoxically, by focussing on the forms of the language in minimal or non-existent

contexts, can claim generalisability to any context or situation precisely because they measure it in none. This would be true if it were possible to focus on 'form' to the total exclusion of 'meaning'; in the absence of any evidence that this is so ('non-existent contexts' are themselves non-existent!) we cannot consider the argument against generalisability to be well-founded.

3.4. Criterion-referenced testing

The fullest account of attempting to define ability in test-taking terms is to be found in the criterion-referenced assessment school. We use the term 'criterion-referenced assessment' to refer to the philosophy of test construction exemplified in the literature, and not to the less radical, though no less important, view that criterion-referencing simply refers to a use of a a norm-referenced test (cf. e.g. Davies 1982b).

3.4.1. Definitions

Black and Dockrell (1980;53) distinguish three types of criterion- referenced tests:

1. Single Act Tests (e.g. measuring a line);
2. Closed Domain Tests (e.g. define something - a matter of sampling, or estimating from a sample);
3. Open Domain Tests (e.g. showing understanding of a concept, or discriminating).

The relationship between domains and attainment could be imagined as a graph whose horizontal axis represents increasingly demanding attainments (e.g. knowledge of specifics, concept attainment, application of knowledge, higher mental processes etc.) and whose vertical axis represents increasingly large domains (e.g. Shylock , *The Merchant of Venice*, Shakespeare plays, plays in general etc.).

The further up the domain scale one moves, the less precise can the diagnosis be; and as a corollary to this, the smaller the domain investigated, the better. Another example of large scale to small scale domains might be the differences involved in testing,say, verbs in French: present tense of verbs in French, the verb *etre*, knowledge of the form *je suis*.

3.5. Diagnostic testing

For Brown (1981;iii) diagnostic testing is simply "one form" of criterion-referenced testing, and as such the two terms are more or less synonymous.

In another view (Davies 1977;48) a diagnostic test is simply a "non- achievement" test, and thus the two terms diagnostic test and achievement test are complementary.

In the latter sense, and referring to L1 reading, Pumfrey (1976) suggests that a [diagnostic] reading test is a means of determining with some precision the extent to which a child has approached one or more goals of a school's reading instruction programme (p.11). He adds that the purpose in testing reading is to provide the teacher with the information that is needed in order to decide the stragegy required to improve the child's reading competencies; thus diagnostic procedures based on reading tests, though imperfect, provide a valuable point of departure from which to further our understanding of the reading process and our ability to help our pupils overcome reading difficulties. "At all levels the diagnosis of reading difficulties is a process of hypothesis generation followed by an intervention, the effects of which lead to a further modification of the hypothesis and thus of the intervention." (Pumfrey 1976;15)

The testing of reading, for Pumfrey, is no more than the careful sampling of some important aspects of a child's behaviour related to reading. He classifies the activity of the testing of reading within three dimensions:

1. Which of the *goals* of the reading programme does the testclaim to measure? Formulated in terms of (a) attainments (reading skills) and (b) attitudes towards the activity.

2. From what kind of *source* is the information collected? Described as (a) informal tests of reading; (b) standardised tests of reading; (c) criterion-referenced tests of reading.

3. What is the level of interpretation to be, i.e. to what *use* will the information collected be put? (a) descriptive; (b)diagnostic - (i) historic; (ii) predictive; (c) evaluative.

Bennett (1974;293) considers that the diagnostic test "measures the distance travelled by the learner and ... determines the point at which the learner went wrong." Bennett is concerned here with foreign language learning and suggests that the value of diagnostic probes is "wholly dependent on an explicit arrangement of items for learning." This we can accept as being in conformity with the criterion-referenced interpretation of diagnostic testing. However, to go on to conclude, as Bennett does,

that "if there is no psycholinguistic theory then there can be no sufficient principles to guide the diagnosis" (loc.cit.) seems to raise other questions and to go beyond the limits of diagnostic testing viewed simply as ,say, non-achievement testing.

Horne (1984;159-161) places diagnosis within the five-fold framework of reasons for testing proposed by Katz (1973):placement, prediction, assessment, diagnosis, evaluation (cf. Davies, 1977: achievement, proficiency, aptitude, diagnostic). The diagnostic function of testing, for Horne, is largely used to determine whether or not an individual is failing. Ideally this diagnosis should also provide indicators as to why this failure has occurred and how it can be remedied.

Horne further draws our attention to a certain ambiguity in the use of the term diagnosis: "... the term diagnosis is here being used as the method(s) by which reasons for learning failure are determined. Nitko and Hsu (1974) note that this definition is not always accepted and that a confusion between placement and diagnosis is often made. The reason for this is that diagnosis is often used for the process of determining to which treatment group a failing pupil should be assigned. Such a process only considers the nature of the failure and not the reasons for that failure."

Diagnostic tests as non-achievement (or criterion-referenced) tests share common ground with achievement tests and as such are closely related to the curriculum, the coursebook, the content of instruction and definition of content domain (and thus have much to do with questions of content validity).

Diagnostic tests as psycholinguistic probes (in the sense used by Bennett 1974) and as instruments for determining the reasons for failure to learn (thus retaining the medical metaphor implied by 'diagnosis') are altogether more problematical and relate to construct validity and, in language testing, the use of 'proficiency tests'.

In this latter sense, Wiener and Cromer (1970) demonstrate six different possible models for conceptualising reading difficulty, all based on antecedent-consequent relationships. They may be summarised as follows:

1. If A, then X
2. If A or B or C ..., then X
3. If A, then X_a, or X_b, or X_c ...
4. If A and/or B and/or C ..., then X_a or X_b or X_c ...

5. If A, then X_a; or if B, then X_b; or if C, then X_c ... If A, then X_a; and if X_a, then B; and if B, then X_b; and if X_b, then C ...

[where A,B,C... are particular and independent antecedents and X_a, X_b, X_c ... are classes of reading difficulty.] (Wiener and Cromer 1970;147-50).

Wiener and Cromer point out that the fourth model above is the most popular form of conceptualising reading difficulty, but that model 5 is the most acceptable form, because the relationships between the antecedents and the consequents are, at least in theory, specifiable. Model 5 assumes that each of the manifestations of reading difficulty is a member of the general class called 'reading difficulty' and that each of these forms is independent. Model 6 conceptualises the manifestations within the class 'reading difficulty' in a model which includes the notion of sequence.

The problem with this approach to diagnostic testing of language would lie in the fact that we have to rely on the notion of hierarchies; if domains are not inclusive and domain orders on the skill continuum cannot be determined, then we cannot set up antecedent-consequent relationships. Horne (1984;164) points out that since we have no methodology for validating domain order "learning hierarchies, in the present state of the art, must not be used as the basis for diagnostic tests nor as the source of test construction theory."

In the context of L2 testing, Hughes (1983;31) points out that it is no use relying on batteries of tests to provide us with diagnostic information on students studying foreign languages; he claims that as (general) measures of language ability batteries of language tests are fine because they possess a high degree of reliability, but if the purpose of the test battery is diagnostic then there is a problem: scores on grammar tests, for instance, have not consistently revealed a grammar component.

3.5.1. Diagnostic testing and CALL

Two approaches to diagnostic testing can be discerned within CALL; the first tries to offer a complete theory, relating diagnostic testing to remedial sequences, while the second uses diagnostic testing more as a means of pre-instructional preparation. These will now be discussed in turn.

3.5.1.1. Diagnostic testing and remedial sequences

This approach is best seen in Ferraris *et al.* (1984), who define diagnostic tests as "formative tests oriented to supply an exhaustive description of student achievement consequent to a learning process." (p.407) They suggest that one reason

for the fact that diagnostic tests are not very diffused in educational contexts is that traditional assessment tools are unable to avoid the enormous student/teacher workload required for the administration of this type of test and for the analysis of results. This is clearly an important point and one which suggests that computerised testing could be of benefit.

For Ferraris *et al.* diagnostic testing is closely linked to programmes of instruction: "... the same diagnostic test can be used both to detect the actual student prerequisites *before* a learning process and to supply information about student achievements *after* a learning process." (p.407)

We see here the intimate connection between diagnostic and achievement testing. Unfortunately, the approach to 'automatized diagnostic testing' outlined by Ferraris *et al.* is based upon three critical assumptions which may not hold for units of language beyond an extremely limited area: firstly, "the subject matter structure used as a basis for testing must be the same as (or analogous to) the structure used as the basis for the instructional processes independently of the media used for delivery." This need not be a serious problem, provided that language teachers' perceptions of what is to be taught do not differ greatly.

Secondly, "given a point of view, it is always possible to represent a given subject in a hierarchical pattern." This seems more contentious; as we shall see, the notion of causal hierarchies must be carefully examined, and it seems fairly certain that to base a theory of testing or instruction on such hierarchies must be misplaced.

Thirdly, "a test is said to be diagnostic if it checks all the student's skills and/or knowledge corresponding to the nodes of the hierarchy." As before, to rely on hierarchical nodes may be a mistake. Within this framework, however, diagnostic testing may be defined as "an instructional procedure suitable for detecting which nodes of a hierarchy the student has achieved and which nodes he has not." The framework will be more useful from our point of view if we ignore the 'hierarchy' and concentrate on the 'nodes'. This becomes clearer if we consider Ferraris *et al.*'s own definition of the node as "... a class of tasks ..." This interpretation provides the basis for an operational definition of 'achieving' a node: "A student has achieved a node N of the content hierarchy if he can accomplish any task of N-task set" (p.409); and it is further assumed that "a student is able to accomplish any task of N if he is able to perform a finite subset of N-task set covering all the nodes of the hierarchy subordinate to N" (ib.)

Thus "for each node of the hierarchy the diagnostic test should include a suitable

number of items which, when answered correctly, insure the achievement of all subordinate nodes." (ib.) This would seem, in effect, to take us into the realms of item domains, to be discussed in chapter 4.

So far as language testing is concerned, 'nodes' are likely to be aspects of the pedagogic grammar or sets of related vocabulary items; but 'nodes' are not going to be related usefully to each other, least of all hierarchically. Performance on a set of items dealing with, for example, the present perfect (or some use of it) does not necessarily tell us anything about ability to manipulate and use other tenses.

Despite an apparent over-reliance on the notion of 'hierarchies', Ferraris *et al.* ultimately reach a position similar to the one we shall advocate here: "The result of a diagnostic test is a table where each node of the content hierarchy is associated with one of these values: 'achieved' or 'failed' " (p.412). Any remedial sequence derived from such a table, or 'profile' is aimed at leading the student from the state where only some nodes are marked as 'achieved' to a state where all the nodes are marked as 'achieved' (p.413). Thus the main idea is that CALL sequence design and diagnostic test construction may take place in an analogous manner.

Ferraris *et al.* themselves admit that if the hierarchy used as a basis for test and teaching sequence construction is a complex one then we have a two-fold problem: (a) methodological, in that it is difficult and often arbitrary torepresent complex knowledge through a rigid structure such as a hierarchy; and (b) technical, in that tests based on a complex hierarchy require large memory storage space and a large amount of time for test structure and strategy definition.

If we view the problem of testing language as a 'shallow' tree with many nodes of equal status rather than as an ever-divisible hierarchical tree then we may gain something from the ideas outlined above. Otherwise not.

3.5.1.2. Pre-instructional diagnostic testing

Pre-instructional diagnostic testing has been discussed by Ariew (1979) and Ariew (1982). Starting from the premise that individualised teaching as a goal for CAI should be axiomatic, Ariew (1979) points out that the kinds of individuation needed in L2 instruction are particularly complex. Moreover no existing placement tests give the precise diagnoses needed to 'interface' with a CAI curriculum: "In a CAI context, an overall score is not sufficient ... With a complete profile of the student's language skills, CAI units can be used to shore up language weaknesses and to assure a smooth transition into the course sequence." (p.331)

The diagnostic test developed by Ariew is designed to evaluate students' ability in French morphology, syntax, audio discrimination, audio comprehension and reading comprehension to an intermediate French level. The test program stores results based on approximately 300 different features of the language "normally encountered during the beginning years of study." It will be noted, therefore, that such a test depends upon pedagogic practice; whether this was intuitive or whether commonly used instructional texts were analysed for common features is not made clear. Ariew provides only one example and it is not immediately obvious to which category this item belongs: the student is asked to answer a question in the affirmative, replacing nouns with pronouns.

"Avez-vous promis ces fleurs?" The desired answer ("Oui, je les ai promises") contains two tested features: the first is the pronominalisation of 'ces fleurs' and the placement of the pronoun 'les' before the auxiliary verb (though it could surely be argued that this in itself is a test of two features). The second feature is the agreement of the past participle with the preceding direct object. If the student answers the question correctly he is awarded two points, one for pronominalisation and one for past participle agreement. As student responses are evaluated a matrix of approximately 300 cells is filled with awarded points for each student taking the test. The cells represent the various features of French (e.g. correct use of the direct object pronoun, adjective formation, adverb formation etc.). The profile thus obtained may be used either to prescribe appropriate CAI materials or to recommend placement. In this case the profile is keyed to a particular course book (Fernand Marty's *Elements for self-expression in French*) and remedial work specifying pages and even paragraphs may be recommended.

It is clearly the case here, though not explicitly stated, that the diagnostic test is built upon this book and therefore we have a good example of the 'specific' versus 'generic' question confronting us again. If a student has not followed Marty's particular method he may be penalised (or rewarded) for something he has met already but in a different form.

On a more general level, Ariew (1982) discusses the problem of computer storage of items used in FL (diagnostic) testing. In an attempt to move away from the apparently specific base for his (1979) diagnostic test Ariew (1982) simply states that storing items by grammatical category is desirable, seemingly because most course-books are constructed in this way. In this system, there are three basic item-types: (a) listening comprehension/ audio-discrimination (b) writing ability, including the ability to handle grammatical problems and (c) reading ability. Because

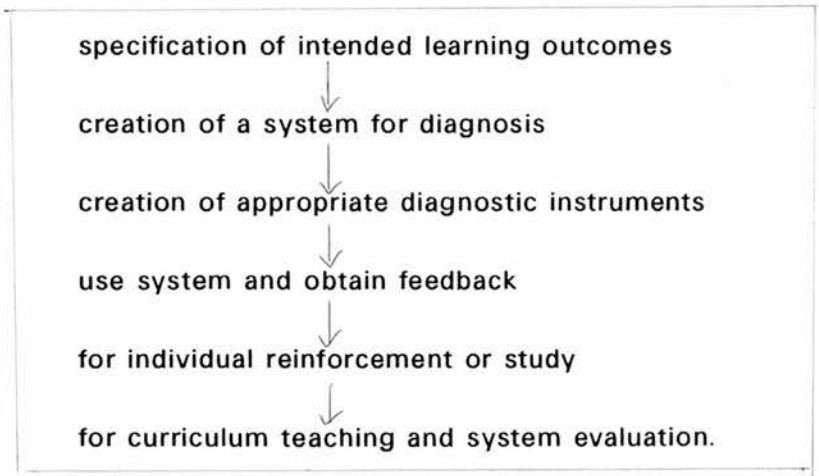
of the storage system 'reading' alsoincludes 'ability to identify and understand certain structures and vocabulary items'; while one may not entirely disagree with this classification, its limitations should be recognised. Furthermore, as with, for example, Comite and Russell's (1982) 'Micro-Deutsch' program, the test user depends on the fact thata soundly constructed pedagogic grammar lies behind the final product.

3.5.2. Diagnostic assessment in practice

The most thorough-going attempt at full-scale diagnostic assessement in education to date has been the Scottish Council for Research in Education's project: Diagnostic Assessment in Secondary Schools (in Geography, Home Economics, Technical Education and Modern Languages initially). In the literature on this project there does not appear to be an attempt to establish hierarchies as such, rather: "Diagnostic assessment is a means by which the teacher and the pupil can find out what a pupil has or has not managed to learn, and is therefore a guide to subsequent action." (Black and Dockrell 1980;1). There is thus a clear statement of the connection between diagnostic assesment and criterion-referenced measurement. In this view diagnostic assessment is not a "revolutionary innovation" but a systematisation of good teaching practice, which requires a specification of unambiguous intended learning outcomes, an absolute criterion against which acceptability can be judged.

Black and Dockrell's model for the creation of diagnostic systems looks as follows (from op.cit.;10; figure 6):

Figure 1
Model for the creation of diagnostic systems



Since diagnostic assessment requires clarity concerning outcomes of learning

intended by the teacher, Black and Dockrell pose four questions to help clarify those outcomes: firstly, can the broad taxonomies of objectives be used in stating intended outcomes? Here there are four levels at which the content of the course is set down in terms of the behaviour which is expected of the pupil who has successfully attained the intended outcomes:

1. The pupil can be expected to recall specific elements of knowledge (expected behaviour is that of 'recall');

2. The pupil can discriminate between examples and non-examples of the concept (=concept attainment; expected behaviour is that of 'discrimination');

3. The pupil is expected to apply the knowledge or concept in new situations (expected behaviour is 'application');

4. The pupil uses a range of elements of knowledge together in problem-solving or analytical situations; this is a 'higher mental processes' category (the modern languages example includes writing, comprehension and listening).

Secondly, can the intended outcomes be grouped according to what it is expected that pupils will be able to do with them? Here a distinction (first drawn by Eisner) is made between 'instructional objectives' and 'expressive objectives' (learning facts *versus* 'encounter/ diversity). In the context of a foreign language this might mean the difference between learning the vocabulary of bull-fighting and writing this in a short descriptive essay (instructional) and learning the vocabulary of an aspect of life in the foreign language which interests the pupil, who then presents it (expressive).

Thirdly, how do these intended outcomes relate to the process of learning and teaching? At this stage Black and Dockrell make the point that the decision on where to test is crucial for diagnostic assessment, and they distinguish three distinct categories of intended outcome related to process:

1. *Modular*, for example, learning the French vocabulary of farming/ understanding the importance of farming to the French economy;

2. *Longitudinal*, for example, learning to use the future tense in French/ developing an understanding of aspects of life in France;

3. *Background*, for example, developing oral/aural skills in French/ developing self-confidence in communicating in a foreign language.

The modular category is unique to a particular unit of work (learned and tested within the unit/module) e.g. specific vocabulary. The longitudinal category is assessed over a longer period; it will be taught intermittently and can be assessed and remediated over a longer time-scale than modular intended outcomes. As for the background category, many background intentions are closely interwoven with the nature of the learning experience (including the development of the skills of discussion, criticism and comparison).

Fourthly, how do these intended outcomes relate to the needs of individual learners? This relates to the issue of 'core' versus 'extension' objectives. For Black and Dockrell the most important function of diagnostic assessment is to allow the teacher to pinpoint the pupil with difficulties in the core (!). Once a pupil has been shown by the diagnostic test to have attained the core, the extension learning outcome serves as a focus for his further learning, at a higher level of complexity or in greater depth. Thus the major focus of most diagnostic assessment will be seen to be modular, with very little distinction in many cases between teaching and testing. For the reason that "we are not interested in how much better a pupil has attained an intended outcome, but whether or not he has attained it" (Black and Dockrell op.cit.;28) the emphasis will be on criterion-referenced tests. Black and Dockrell propose the following three features of diagnostic test design:

1. Establish individual intended outcomes - a clear specification of what is to be assessed is needed (this reflects Popham's *general description*);

2. Information is required on discrete intended outcomes rather than general attainment; diagnostic items should be testing oly one thing at a time and should presume as little as possible about factual knowledge unrelated to what is being tested;

3. It is important to sample a pupil's understanding by using a set of items all testing the same sectioin (or domain) of knowledge. Here we have again the problem of domain inclusivity.

In the end, Black and Dockrell emphasise the following points:

1. A reliable criterion-referenced test used for diagnostic assessment will comprise groups of items, each group sampling only one domain.

2. Each item will be carefully written to minimise ambiguity and will assume as little as possible of the pupil's ability to perform on skills .

3. The criterion score will accomodate the possibility of individual

pupils making random errors.

4. There will be a number of items testing each domain.

3.6. Conclusion

Concentrating too much on criterion-referenced testing and content analysis can be counter-productive: Wood (1976) has put it thus: "One of the features of criterion-referenced testing I find hard to stomach is its *exhaustiveness* when in practice there has to be selection, usually severe."

However, a positive approach to criterion-referencing as a method of test construction embraces all the important elements which need to be taken into account. As Haertel (1985; 34-35) points out, a complete manual for a criterion-referenced test must include description of achievement constructs (i.e. prior validation of the testing domain) as well as empirical validation on actual test data. Thus the test writer has to produce a description not only of the content specifications but also some characterisation of cognitive processes or memory structures (thus calling on the insights of curriculum specialists and educational psychologists). The curriculum content and psychological process descriptions would have to be augmented by the specification of a faceted domain (see Chapter 4) of behavioural outcomes implied by the construct.

A major difference between this procedure and conventional criterion-referenced testing (or norm-referenced testing for that matter) would be in the use of domain descriptions. Current practice is to specify a content domain of potential test items as precisely as possible and to sample that domain to build a test. The domain specification is critical because it is the unique operational definition of what is measured. The procedure specified here would employ two domain descriptions: the first would be a faceted domain of behavioural outcomes (arising from considerations of the construct under consideration, such as we have explored in Chapter 2), and the second would consist of possible test items (which is what the process of item bank design represents).

How domains might be specified is considered in the next chapter.

CHAPTER 4

ANALYSING THE 'DIFFICULTY' OF READING TEST ITEMS

4.1. Introduction

An analysis of the difficulty of reading test items is closely connected with the analysis of 'attainment' or 'ability' as discussed in Chapter 3; Pollitt *et al.*'s (1985) investigation into the difficulty of Scottish O-grade questions "forced us to clarify the distinction between 'intelligence' and 'attainment'" (op. cit.; 5). The problem is that the concept of attainment is rarely discussed explicitly, but rather has to be inferred from the syllabus laid down, from the topics included in examinations and from the types of questions asked. Questions may make very different intellectual demands on a candidate; often the form of the question may have been specifically chosen to make particular demands using some classification such as the familiar Bloom taxonomy ('knowledge, comprehension, application, analysis, synthesis and evaluation'). 'Intelligence' is typically defined in terms of intellectual skills which are intended to be 'content free'; a typical 'intelligence' test will demand from the candidate the ability to undertake logical analysis at speed, to grasp relationships quickly, to spot absurdities, illogicalities or *non sequiturs*, to make valid generalisations, and to think effectively in terms of abstract concepts where necessary (Pollitt *et al.* 1985; 5). How does this differ from the way we tend to think of language tests of reading comprehension? More importantly, how can we ensure that our test questions test 'attainment' rather than intelligence?

4.2. Test task

4.2.1. Effect of test task on the reader

4.2.1.1. Introduction

Fillmore's (1982) observation that a humorous passage, if presented a segment at a time, will lose its humorous nature, leads one to wonder just how far the reader can be affected by the fact that any testing situation necessarily involves a degree of artificiality and may therefore lack validity as a measure of the testee's reading skill.

More specifically, how do the demands of the test task affect the reader's processing of the test task? Royer et al. (1984) suggest that the level of detail learned from a text will vary depending on what the reader wants to learn from the text: reading intent can also affect the very nature of the information that is acquired from text. They set out to determine whether the reader's intent can be brought under

experimental control by manipulating task demands while reading. In particular, they ask: "Is it possible to manipulate and control intent so that readers will either learn more from a text, or derive something from a text that is qualitatively different from what they would normally acquire? (Royer et al. op.cit.;66).

Three ways in which reader intent might be altered were considered:

1. Through learning objectives;
2. Through inserted questions;
3. Through higher order questions.

We shall now consider these in turn.

4.2.1.2. Learning Objectives

Duchastel and Merrill (1973) found that five studies demonstrated superior test performance for subjects who received learning objectives, but that five other studies demonstrated no effects, either positive or negative. However, Duell (1974) found that the learning objectives group significantly outperformed the group without objectives on those test items that had been judged by the subjects to be unimportant (i.e. names, and dates associated with principles discussed in the passage *versus* definitions of principles and applications of principles to new situations.) Rothkopf and Kaplan (1972) (replicated by Kaplan 1976 and Rothkopf 1974) found that a group given specific objectives outperformed agroup given general objectives, which was not due to a contrast between lower-order and higher-order skills/objectives.

Royer et al. (op.cit.;69) conclude that objectives have been shown to enhance the amount learned from a text when learning is measured by performance in test questions related directly to those objectives; otherwise, there seems to be an inconsistency in results.

4.2.1.3. Inserted questions

Rothkopf (1966) found that intentional learning was greater for all groups that had been provided with the inserted questions than it was for the control group. His interpretation of these results was:

1. Inserted questions, whether pre- or post-, facilitate the learning of question- specific information by directing attention to relevant parts of the text;

2. Inserted post-questions have an additional facilitative effect on the acquisition of general skills related to the inspection of to-be-learned material, resulting in increased incidental learning;

3. Providing subjects with answers probably lessens the attention to the text during reading.

4.2.1.4. Higher order questions

Royer et al. (op.cit.;74) comment that it seems reasonable to assume that questions which focus student attention on relatively superficial aspects of some information should lead to a different kind of understanding than would questions directing attention to more complex aspects of the same information. However, in spite of the "long history" of the idea that the nature of the questions asked affects the level of learning, there is little firm evidence one way or another. Royer et al. conclude that many of the experimental results may be due to a 'laboratory effect'. Alderson and Urquhart (1984;85) suggest that many of the results depend on the compliance of the subjects and that it may be that as long as the student considers the material to be irrelevant, he will accept the teacher's formulation of the goals. This is clearly usually the case in the testing environment, and therefore suggests that we need not worry too much about the 'communicational relevance' of our test material.

4.2.1.5. The differences caused by the testing method

Shohamy (1984) has studied the question of whether the testing method makes a difference in the case of reading comprehension in English as a foreign language and in Hebrew as a first language. Multiple choice and open-ended testing methods were used for the L1 (Hebrew) and L2 (English), which resulted in four testing methods in total. Shohamy developed a two-part test, the first part of which included eight short texts each followed by one multiple choice question directed to the main idea of the text (or occasionally to a vocabulary or grammar item which was somehow thought to be instrumental for the comprehension of the text); the second part was a longer text followed by eight questions also directed to the main idea as well as to vocabulary and grammar items.

The results showed that the multiple choice Hebrew (L1) version was the 'easiest' text, while the open-ended English (L2) version was the most 'difficult'. Shohamy concludes that while multiple choice items are 'easier' than open-ended items and Hebrew is 'easier' than English (for this sample), in fact low-level students are more sensitive to the testing method and text, while the high-level students are hardly affected by these variables.

A further source of difficulty in this study was the language of the questions: presenting the questions in L2 introduced an 'unnecessary' source of difficulty.

While there is limited evidence here for the view that the testing method affects scores on reading comprehension tests, the kinds of differences here reported seem to be coarsely distinguished; what seems to create the 'difficulty' is whether the testee has merely to respond or whether he has to construct a response. This seems entirely in keeping with intuitions about what is difficult in language generally, but fails to address the more subtle questions of which *types* of multiple choice questions, if any, are more 'difficult'. Questions of this latter type have been partly addressed by Bensoussan *et al.* (1984) who set out to examine to what extent and by what means the ease or difficulty of multiple choice questions (in tests of EFL reading comprehension) could be altered by the test constructor. The following three factors were identified initially as being possibly responsible for affecting the difficulty of multiple choice questions on a reading comprehension passage:

1. Changes in distractors. For example, true statements which do not correctly answer the stem question, or greater homogeneity of distractors would probably create harder items; wrong paraphrases with delicate distinctions would also create difficulty. On the other hand, using simpler language, easier paraphrases, using information not found in the text itself, or wrongly paraphrasing parts of the text in an obvious way would probably create easier items.

2. Using completely different questions. It is not entirely clear what Bensoussan *et al.* have in mind here; probably they refer to the use of a wide variety of question types tapping supposedly different reading skills – a 'main idea' question followed by a 'vocabulary in context' question perhaps.

3. Local or global level of questions. The need to understand a word, phrase or sentence in the text would probably result in an easier item than the need to comprehend two or more sentences, a paragraph, or the whole text. Since a text contains more words and phrases (local level information) than sentences and paragraphs (global level information), Bensoussan *et al.* reasoned that local text-level questions, focussing on specific linguistic information, would lend themselves better to adjustment of difficulty level than global text-level questions dealing with more general ideas which appear in longer stretches of text.

However, the hypothesis that the test with the putatively more challenging set of questions would be more difficult was not supported. The only conlusion that Bensoussan *et al.* felt able to draw was that there appears to be some tendency for local text-level questions to be more readily affected by linguistic changes than global text-level questions.

The problem really is that we do not have any criteria for deciding what is likely to be a 'difficult' item type, beyond a certain intuition based on vague ideas of what reading in the L1 is like. Perkins and Jones (1985) claim to have found that items which test "complex inferential skills - drawing conclusions" are high in difficulty (as are a variety of others), but the study upon which these conclusions are based seems to be flawed (see section 4.2.3.2. below).

4.2.2. Text type and testee motivation

Other constraints on the validity of tests of reading comprehension concern the student's motivation and attitude.

Doerr (1980), working with adult ESL learners, used two 100-word passages arguing for and against three controversial topics and tested students through cloze and oral interview. She points out (op.cit.;135) that attitudes have significant effects on native speakers , who learn 'covaluant' material (material they tend to agree with) more effectively than they do contravaluant material, and that "we might expect the effects to be more pronounced in learners who are not as sensitive to the redundant features of the language used."

However, on the basis of her findings, Doerr (op.cit.;137) concludes that it is not possible to argue strongly that a foreign language learner's self-expressed attitude toward a controversial issue has no relationship to his comprehension of a contravaluant message concerning that issue. However, the effect is not substantial enough to override the slight difference in difficulty across the *pro* and *con* texts. In other words, the hypothesis that the foreign language learner will tend to block out or alter contravaluant material had to be cautiously rejected.

The question of motivation in approaches to learning and test performance has been investigated by Fransson (1984), who discovered that a subject motivated by expected test demands to read a text for which he has very limited interest is likely to adopt a surface-learning strategy, while deep-level learning seems to be the normal strategy chosen by a student motivated only by the relevance of the content of the text to his personal needs and interests.

Alderson and Urquhart (1984;120) comment that subjects who did not adapt to expected test demands were overwhelmingly 'deep' processors, while subjects anticipating and hence adapting to a test set by the experimenter were far more inclined to be surface processors.

Guzzetti (1984), on the other hand, using a miscue analysis with native speaker subjects, conducted a study designed to test the idea that regardless of the content of the reading material, readers made similar uses of reading strategies to gain meaning. Guzzetti concluded that high, average and low ability level readers are consistent in their use of syntactic and semantic cues to reconstruct meaning: "the application of these strategies does not vary with the content of the reading material." (Guzzetti op.cit.;660) The data of this study do not support the position that specialised skills are needed for reading particular content areas; on the other hand, students' personal interest and knowledge in a content area can affect their reading performance in that content area, a fact which is said to support the view that internal cognitive schemata, which reflect personal interests and experiences, enable readers to reconstruct a written message (Guzzetti op.cit.;666).

On the other hand, Alderson and Urquhart (1983), using a cloze test with L2 subjects and investigating the assumptions underlying 'general interest' texts in tests of ESL proficiency, concluded that there was support for the hypothesis that students from a particular discipline would perform better on tests based on texts from their own subject discipline than would students from other disciplines. That is to say, students appear to be advantaged by taking a test on a text in a familiar content area (Alderson and Urquhart 1983;126).

Results, then, are inconclusive.

4.2.3. The role of factual knowledge and passage dependency

4.2.3.1. Information gain

One strand that emerges from the foregoing discussion is the extent to which perceptions of the uses of comprehension tests differ; for some, a test is a measure of how much a testee can learn from a particular passage, for others it is a test of language manipulation.

Anderson's (1972) criticism of comprehension tests relates crucially to the test as a test of what has been learned. Anderson suggests that comprehension tests do not assess what a reader has comprehended from a passage in terms of new information, but rather measure a much more general overall language competence. Anderson maintains that his "verbatim" and "transformed verbatim" question-types cannot really be said to test the reader's comprehension at all. He sees the task of the comprehension test instructor as the devising of questions which can be answered if a person has semantically encoded the meanings of the text and assimilated them as

new information.

Harrison and Dolan (1979;18–19) comment on this view and suggest that a contrast be set up between 'comprehension' and 'information gain'; the latter could be measured by, for example, giving testees questions before and after reading a test passage to see how much more they 'knew' as a result of reading.

Information gain techniques seem to have the effect of cancelling out the effects of overall language competence; but testing reading comprehension in L2 demands a shift from this inflexible position, say Harrison and Dolan (loc.cit.), because the purposes of comprehension testing in L2 seem to be rather broader than those in L1. It may be, for example, that the content of the passage, were it in L1, would be immediately comprehensible to the reader. His problems in comprehension may be wholly related to aspects of grammar and syntax, which are trivial for the L1 user. Thus a comprehension question which would be testing the minimal level of reading ability in L1 might conceivably be worth posing for certain L2 language users.

4.2.3.2. Passage dependency

Tuinman (1974) examines a related problem to the above: the extent to which questions used in the test of reading comprehension could be answered without reading the passage upon which those questions are based. Tuinman points out that tests of reading comprehension puport to measure how well a student understands what he is reading and that the questions used to ascertain the degree of this understanding are based on the tacit assumption that a direct relationship exists between reading a passage and answering questions about it. But in the case of a great many reading test items from standardised tests "this is a faulty assumption" (op.cit.;208).

Lack of passage dependency signals potential invalidity more than actual lack of validity; only if an item is responded to without prior reading of the paragraph to which it refers does that item constitute an invalid measurement in the context of a reading comprehension test. For this reason low passage dependency is only a threat to valid measurement and not proof of invalidity. (Tuinman op.cit.;211)

Using data from five major standardised tests, Tuinman found that three of the tests allowed a student who did not have access to the passages to obtain a score as high as 70% of that of a student with the passages. On the average, for these tests, not reading the passage resulted in a loss of performance of less than 30%.

Inference items are clearly more passage dependent than factual items. (cf. also Slade and Dewey 1983: the role of grammatical clues in multiple choice questions)

Perkins and Jones (1985) have attempted to investigate the question of how far there is passage dependence in tests of EFL. They began from the assumption that foreign students admitted to full-time study as undergraduates in the States can read and write the English language on a par with native speakers - this seems unduly optimistic. Further, the undergraduates were tested on two L1 reading tests designed for high school students. The results of the study showed that the passage did not contribute very much to the reading process - in other words "the majority of items were assessing background knowledge which was not gleaned in the reading process" (op.cit.; 147). On the other hand, because "many, if not most, of the scores were clustered near the zero point of the continuum" it seems more likely to conclude that the language of the test/text was simply too difficult for this particular population.

4.2.4. Discourse structure

4.2.4.1. Skills or general cognition?

As discussed in Chapter 2, there is considerable difficulty in distinguishing between language specific skills and aspects of more general cognition.This should be borne in mind in the discussion which follows.

4.2.4.2. Skills defined as tasks

It is not necessary to accept the existence of 'reading skills' in order to use test questions which are described as 'testing reading skills'. Lunzer and Gardner (1979;68) suggest that different skills are 'comprehension tasks': "they describe the sort of questions that one can and should include in a varied and interesting comprehension test ..[which is] an indirect measure of the adequacy of reading."

4.2.4.3. Text and task difficulty

Lunzer and Gardner (loc.cit.) suggest that the relative difficulty of a question or of an interpretation is mainly due to the difficulty of the text on which it bears: "it is not due to some hypothetically distinct differences in the thinking process associated with the question type."

This, however, would seem to be inadequate if it is suggesting that question difficulty has no bearing on the process of testing comprehension. Bormuth (1969;52) points out that, while comprehension is a response to the language system and not

just a set of mental processes which can be defined independently of language, in using a test there are at least six different sources of difficulty for the testee:

1. He has to read the language stimulus;
2. He has to comprehend the language features of that stimulus;
3. He has to read the test task;
4. He has to comprehend the test task;
5. He has to derive an answer to the test task;
6. He has to answer the question.

Brown (1983) has similarly suggested that for listening comprehension, understanding can break down at one of three points: the language stimulus, the test task, the processes going on in the individual testee.

Bormuth (1970;6) further suggests that the failure to achieve operational definitions of test items means that in the final analysis a test item bears a certain label just because the test writer and his associates say that that is what it measures. The point, for Bormuth, is that traditional labels such as 'evaluation' or 'comprehension' refer to mental processes and not to observable events, so when the test writer selects such a lable, he is using it to refer to something which occurs only in his private mental life: "it is highly questionable whether the same labels mean the same things to two different test writers." (op.cit.;11)

4.2.4.4. Discourse cloze techniques

Recent attempts have been made to test understanding of reading comprehension through the use of cloze techniques, which are said to go beyond the level of sentence comprehension. In particular, the framework of discourse analysis has been used to provide the theoretical justification for such attempts, resulting in a 'discourse cloze' test. (cf. e.g. ELTS G1 section 32)

Levenston et al. (1984) describe the use of cloze techniques to test reading comprehension through discourse analysis. They claim that just as a text can be viewed as a hierarchy of units at different levels of analysis - from word to sentence to paragraph to discourse - so understanding a given text can be seen as a hierarchy of skills corresponding to these levels:

- recognising the content words

- knowing the semantic function of grammatical structures
- construing the sentences
- identifying the inter-sentence relationships
- supplying an overall interpretation for the text in all its interactions
- appreciating the tone in which the whole discourse is written

All of this they term the 'linguistic' component of reading comprehension Levenston et al. op.cit.;202)

It should be clear from what has been discussed in chapter 2 that to identify linguistic elements beyond the sentence is a dubious activity, if this is meant to imply that there is a 'discourse structure' corresponding to syntactic structure (cf. e.g. Morgan and Sellner 1980; Morgan 1981). The controversy relating to the hierarchical ordering of skills has also been outlined. And yet Levenston et al. (op.cit.;202) take the following premise as the cornerstone of their theory: "If the skills are hierarchically ordered then testing the higher order skills, identifying inter-sentence relationships, grasping conveyed meanings, appreciating tone, inevitably tests lower order skills as well; without a grasp of at least part of the word meaning and sentence syntax, one cannot trace the thread of discourse."

Levenston et al. place great stress on 'cohesive ties', which are said to differ from linguistic and pragmatic knowledge in that knowledge of cohesive ties is "specific to text-processing as distinct from sentence processing" (op.cit.;206); the discussion of this question in chapter 2 showed how dangerous it is to make this claim.

What Levenston et al. suggest is a discourse completion exercise in which all items deleted will test this textual component; only those items are deleted which mark in one way or another relationships between propositions. Unfortunately only one example is given of what might be intended by this, so it is difficult to judge the criteria for deletion with any objectivity. It is said, however, that markers of co-reference and connectives between propositions will be deleted; completing the blank for the former obliges the student to discover the antecedent, in other words to identify the relationship between topics of the discourse. The semantic load carried by such an item is said not to be independent but to depend on another element to which it refers. In discourse cloze this second element is located in the text. (Levenston et al. op.cit.;208). This is again to ignore the fact noted by Webber (1980) and Morgan and Sellner (1980) that cohesion is not a property of linguistic elements in the text, but something that readers impose on the text.

The second category of cohesion markers identified by Levenston et al. includes "inter-propositional connectives". Such markers indicate logical and rhetorical functions: temporal sequence, cause, contrast, addition, itemisation, exemplification and so on. In order to restore correctly deletions of this category, the respondent must determine the underlying coherence relationships and find the appropriate surface exponent. It is not clear how this involves anything other than the choice of an item of vocabulary, at least in any way that differs from 'normal' cloze.

Levenston et al. say that their technique is useful in isolating the grasp of cohesion for separate testing (and that a grasp of cohesion is a a factor in reading comprehension of continuous discourse); but from the paucity of examples given it is difficult to see how this 'discourse' cloze differs in any useful way from 'traditional' cloze.

4.2.4.5. 'Authentic' discourse cloze

In an attempt to improve on this, Deyes (1984) offers the framework for an 'authentic' discourse cloze. He criticises Levenston et al. because the limitation of deletion items to cohesive features makes such products 'text-cloze tests' rather than 'discourse cloze tests'. If a truly discourse cloze is to reflect the reader's ability to follow information through the text and use contextual clues as well as co-textual ones, then *theme* and *rheme*, as units of information, provide criteria for item deletion: "Relevance of these units to the comprehension of the discourse can be determined by applying the concepts of 'frame' and macro-structure criteria" (Deyes 1984;128).

Deyes points out that while tests of the type outlined by Levenston et al. ensure that the testee is required to derive his or her gap-filling items from clues beyond the immediate clause boundary, by using the notion of textual cohesion as the principal criterion for item deletion, nevertheless the drawbacks of such an approach are twofold: first, a score obtained by this method is difficult to interpret - it may mean that the student has not understood the cohesive relationship, or it may mean that he has understood it but is not sufficiently familiar with the distribution of items in the system Deyes gives the example of the semantically equivalent interchange of *it* and *this*, where the system demands one or the other but not either).

Secondly, if deleted items are limited to those drawn from the cohesive system we remain at the level of text cloze rather than discourse cloze. That is to say, we are testing knowledge of the language system and not requiring the learners to demonstrate understanding of the communication as a whole. This is similar to the comments made earlier about inter-propositional connectives.

Deyes' central thesis is that interpretation can be tested by requiring students to replace not single words but communicative units (cf. also Bowker 1984). Such units may be recoverable textually, but there should also be some whose replacement shows an interpretation of the wider context in the same way that we sometimes show our understanding of what somebody is saying by completing the sentences for him (Deyes op.cit.;129). Such a system will clearly be more demanding of marker time than other systems.

One of the problems with this technique is the identification of recoverable communicative units. Deyes suggests that recoverable rhemes are those whose content can be derived by knowledge of linguistic stereo-types ("collocations") and/or by stereo-type knowledge about the world. An additional question here is that the difference between stereo-typed and more specific knowledge is likely to vary according to the population being tested. "Given, however, that such degrees of knowledge determine also the native speaker's ability to understand an L1 text, some dependency on the reader's world knowledge in no way seems an unfair criterion to introduce in tests of reading comprehension." (Deyes op.cit.;132)

Deyes' analysis of determining relevant communicative items in a text has affinities with Bormuth's attempts at operationally defining higher levels of discourse, and it is conceivable that some sort of progress could be made here. Deyes infers three levels of propositional importance:

1. Propositions denoting an accidental property of a discoursereferent, where replacement of communicative units would not necessarily be relevant to comprehension of the discourse;

2. Propositions which represent a normal condition, component or consequence of a fact denoted by another proposition;

3. Propositions which define the immediate "superconcept of the micro-position".

Deyes suggests that deletions from the second class of proposition are preferable as the best test of understanding of those major propositions and hence of the discourse as a whole (op.cit.;133), but the idea has yet to be tried out.

4.3. Classifying 'difficulty'

4.3.1. Textual difficulty and readability

The time-honoured way of determining the 'difficulty' of a text has been to use one of the many available readability formulae. There are, however, several problems with this approach.

Stokes (1978) has shown that while the majority of studies investigating the validity and reliability of readability formulae have been concerned with the rank ordering of texts, in fact if we look at the grade levels to which such formulae would assign texts then the mean levels predicted by the formulae are significantly different overall. Moreover, no formula consistently predicts a similar grade level to any other formula.

The main problem is that readability formulae aim to predict and quantify the comprehensibility of a text for its intended readership taking into account such variables as average word length, number of polysyllabic words per *n* sentences or monosyllables per 100 words. But since there is no obvious limit to the number of variables which can be used, the choice of these variables is entirely arbitrary. Readability formulae can only rank order materials, that is, compare them on the same linguistic variables; if there is any validity to this procedure it is to the extent to which there is agreement with existing standards, and as Manzo (1970) has said, this is incestuous and makes readability research a construct without a point of reference.

So far as the readability of foreign language materials is concerned, perhaps the main attempt at arriving at some sort of measure of readability has been the use of T-units (Hunt 1966 and 1971 and Larsen-Freeman and Strom 1977). This is essentially a measure of syntactic maturity as expressed in syntactic complexity; but it is not at all clear how far a simple count of sentence length does not do the same job (R. Baker 1982). Laroche (1979) has suggested that readability measurements for foreign language materials should concentrate on 'linguistic variables' (cognate count, cognate frequency, sentence length, phrase-structure complexity), as in fact is the case with L1 materials; he suggests that readability formulae should be developed for L2 in the same way that they have been developed for L1. This, however, ignores the criticisms of readability formulae in general, and is possibly a step backwards.

Of course, it should not be assumed that readability formulae *define* readability; they merely reflect it. The use of any objective measure of textual difficulty, however, presupposes that the text already exists; it assumes that a writer chose a topic, made

decisions about how to order the ideas within the topic, and then decided to express the ideas in words (Davison and Kantor 1982). There is no room for any 'interactive' view of reading here. It will be noted that the criticism here depends partly on the psycholinguistic arguments as to what readers do when they read a text (cf. Johnston 1983; 21); one thing that formulae cannot distinguish, for example, is a well-written challenging text from a badly constructed text which makes interpretation difficult. Nor can they take into account 'information density'.

4.3.2. Item difficulty

How might the 'difficulty' of reading test items be classified? Pollitt *et al.* (1985; 17) classify all test items initially into one of the following four categories:

1. DH: 'difficult' questions [i.e. questions with a high *difficulty index*] which discriminated well within the sample (H = high)
2. EH: 'easy' questions which discriminated well within the sample
3. DL: 'difficult' questions which did not discriminate well within the sample (L = low)
4. EL: 'easy' questions which did not discriminate well within the sample

These purely statistical criteria were used as the basis for analysing the content 'difficulty' of test items in a variety of school subjects; here we shall be concerned only with their analysis of comprehension questions in English and French.

The first point to be made is that in the Pollitt *et al.* analysis, all the French and English items were classified together. This was because although the investigators had started with the assumption that candidates reading and using their native language, and at this level quite competently, would face somewhat different problems from those required to operate partly in a second language with which they were unfamiliar, yet it became clear as the analysis proceeded that many of the sources of difficulty were common to both subjects, and differed more in degree than in nature. In other words, a number of the characteristics of items which appeared to cause difficulty were general 'written interpretation' skills rather than specifically 'French' or 'English' skills.

Four main sources of error were identified (op.cit.; 55):

1. In reading the question rubric and understanding the task;

2. In finding the correct piece(s) of text, from which an answer might be derived;

3. In understanding the meaning of the identified text piece(s) at the level either of decoding or interpreting;

4. In the composition of an adequate written response.

It was also found that these broad areas covering the process of answering a question were dependent each upon the next, so that the process could be seen as a chronological one, at any point of which a candidate might both drop marks and ruin his chances of completing the rest of the task successfully.

Reading the question is a stage which can be presented to the candidate in a variety of forms: statement leaders could be used to give information about the task, instructions could be given telling the candidate to do something, interrogatives could ask the candidate about the text or ask for an opinion. Each of these parts could be more or less complex in its own right, or might, by its relationship with the other parts, add to the complexity of the question as a whole. The correct definition of the task requires the candidate to synthesise the parts in order to delimit 'outcome space'. Where the individual parts of the rubric of the question contain a number of 'supports' to candidates, outcome space is clearly defined; hurdles on the other hand can be seen as blurring the edges of the definition, forcing the candidate to find his own limits.

The construction of the overall *set* of questions by the setter may be important to candidates if it tends to reflect the structure of the passage; the test-wise candidate may be able to reconstruct the gist of the passage from the question sequence and therefore be searching more systematically for certain bits of likely information than the candidate who approaches each question in isolation. This ability to make inferences across a set of questions is certainly a test of attainment in English, but will tend to favour those candidates who have been taught to regard the questions as an information resource.

Finding the text, having one decided upon the nature of the task, involves recognition of sets of matching 'markers' in question rubric and text, as well as identification of text piece(s) relevant to the task. Hurdles in the way of identification of precise text piece(s) might arise out of the existence of two or more text pieces for consideration, where the correct choice depended on a good match of text to task.

Interpreting the text or 'understanding' text seemed to separate into two distinct

tasks: decoding separately identifiable pieces of text, and, where necessary, contextualising them and processing them in order to formulate an answer. In the Scottish study, 'content' words in a phrase or sentence (meaning-bearers such as nouns, verb stems and adjectives) were taken at face value at the expense of functional details; detail was not read carefully enough. *This was particularly noticeable in French*, ignorance of the meaning of a particular word clearly affected the decoding process, as did the complexity of sentences, particularly where phrases or clauses were embedded or inverted.

Three particular circumstances were found to occur consistently in difficult questions:

1. They required subject specific techniques which are generally taught, such as understanding metaphorical language or the effect of conventions punctuation;

2. They required candidates to follow the time sequence of a passage;

3. They reuqired candidates to 'match' particular words and expressions to perceived meaning and comment on their effectiveness in context, reacting to linguistic subtleties beyond the simple expression of meaning.

Composing an answer need not concern us here, especially since many of the identifiable errors were attributable to the stages in tackling the questions outlined above.

In conclusion, it can be said that there are several dimensions to 'difficulty' in questions in written comprehension tests. The intrinsic difficulty of the words, structures and linkage of ideas and concepts in the passage has traditionally been considered as the chief determinant of the difficulty of the test, and most closely linked with the concept of 'attainment' in whatever language is being tested. However, it is clear that some of the variables identified, concerned as they are with the wording of questions, their syntax and their relationship with the text, are those which control the candidate's access both to the task and to the meaning of the passage, by providing him with supports or putting hurdles in his way.

On the other hand, while all the variables it is possible to identify (with the possible exception of ambiguity in question wording) may have some linguistic validity, over-dependence on variables which are not specific to the instrinsic difficulty of text is clearly undesirable, and will result in assessment of skills which are only

peripherally connected to attainment in comprehension of whatever specific language we are interested in.

We now consider how items might more systematically be constructed to eliminate some of these extrinsic sources of difficulty.

4.4. The Technology of Achievement Test Construction

4.4.1. Definition of 'technology'

'Technology' as used in the testing literature is closely connected with the criterion-referenced movement led by Popham and subsequently formalised, notably , by Roid and Haladyna (1982). It is a term which is perhaps a little unfortunate for its mechanistic associations, but it is this aspect which represents its strength: it is an effort to define test items and content domain more rigorously than before. At the extreme end of the scale it is envisaged that a technology of test-item writing could be handled automatically, by computer say. An emphasis on the technology of item writing serves to clarify the following points:

1. It aims to enable different item writers to produce similar items (cf. Popham 1978), and so to remove the intuitive and unreliable aspects from test item writing. This can be a major problem in that if asked, for example, independently to write items to test, say, 'understanding the main idea' of one passage, different writers will produce totally different questions (cf. Zuck and Zuck 1984)

2. Conversely, it aims to ensure that test items can be clearly seen to be testing what they claim to test.

4.4.2. Testing technologies to date

As used in this section, 'technology' refers to any systematic attempt to relate item-type to content domain.

4.4.2.1. Form and function

It is probably fair to say that most existing theoretical discussions of the L2 testing process tend to concentrate on form rather than function of test items (Harris, Heaton, Valette notably), or at least the form is seen as the focus of the discussion while function is worked out as an afterthought. Thus discussions of the merits of integrative versus discrete point tests have highlighted the problem of what is being measured once the form has been fixed. Within the context of L2 reading Widdowson (1978;95) recognises four types of question:

A. Wh-

B. Polar

C. Truth Assessment

D. Multiple choice.

Although his concern is with questions for their use in developing reading comprehension rather than testing it, it would still seem that this typology is misleading. Why, for example, should multiple choice questions belong in a qualitatively different category to the others? Surely Wh- type questions could have multiple choice answers; and are not types B and C above merely multiple choice questions with only two possible answers?

Widdowson argues that A and B type questions "... suggest a social interaction which does not in fact take place ... C and D type questions ... are context free in that they do not require the learner to take on the role of answerer in a question and answer exchange." (op.cit.;96-97). But *any* interference in text processing by means of questions must "insome degree be an imposition on the learner" and it would be difficult to sustain an argument that C and D type questions are to be preferred, since they are merely techniques for avoiding direct questions, and therefore, perhaps, A and B type questions at one further remove from reality. The point, however, is that we are primarily concerned with form here, and have little to guide us when it comes to filling in the content of the questions.

4.4.2.2. Product and process

What Widdowson seems to be after is some form of process measure. So far as testing is concerned this is probably ill-conceived. Johnston (1983;61) has pointed out that process measures can sometimes distort the reading process by introducing an element which interacts with the ongoing process. This seems unarguable, though as a teaching principle it may have its value. Process measures also include cloze tasks, but unfortunately these do not tell us why readers do what they do, thus providing little diagnostic information, and, as Johnston suggests, place quite a demand on short term memory for what is, essentially, a problem-solving exercise.

The problem of interaction between task and reading can never fully be removed even with product measures; multiple choice questions, for example, still suffer from the problem which, for Johnston, afflicts all forms of probed recall: that the cueing can induce processing which would not otherwise have occurred (Johnston1983;59).

Johnston's suggestion is that we supply a variety of alternatives which suggest different processing strategies so that respondents "no longer find unequivocal incentives toward specific extra processing." (ibid.)

Similarly, inferences which a good reader might make while reading may not be made by a poor reader until the probe suggests the value of making such an inference. This inevitably complicates the interpretation of any probe- type question, at least if we are concerned with the reader's quality of mind.

4.4.2.3. Testing L2 through reading

Concern with language testing through *the medium of print* leads to the clearest attempts to systematise test construction.

Davies (1977) provides a framework within which each of the four language skills can be used to test some aspect of the language system; thus reading can be used to test:

- Phonology (e.g. phoneme rhymes)
- Grammar (e.g. sentence completion)
- Lexis (e.g. synonyms)
- Context (e.g. modified cloze passages or questions on a text)
- Extralinguistic features (e.g. reading speed)

(pp.79-85)

This is fine as far as it goes, but represents little more than a conceptual framework within which to begin the real task of creating items; all questions about test content are avoided (one man's grammar may be another man 's context), and in particular the problem of defining comprehension ('context') tasks is avoided. However it does represent an attempt at defining content domain.

The Cambridge Examinations in English (1982) divide reading comprehension into two parts: the first part "systematically tests usage"; of 25 items 8 will test semantic sets and collocations, 8 will test use of grammatical rules and constraints, and 9 will test synonyms/antonyms, 'semantic precision', phrasal verbs etc. This again appears to be fairly rigid, but given these specifications item writers are going to differ enormously in the type of question they produce. Once more, however, it does represent a useful first step towards content definition.

An extremely good example of principled L2 test construction is to be found in Black and Dockrell (1980), where a test of German datives is given. This is discussed more fully in the next section. The point to be made here is that a fully diagnostic test of even this one small part of German grammar requires pedagogical insight and principled analysis to ensure that information obtained from the test is to be of value.

4.4.2.4. Testing L2 reading

Existing technologies for testing L2 reading (in its own right, as a separate skill area) tend to be extremely loose:

The Cambridge Examinations part 2 proposes that "items will test an understanding of context", which is seen as the testing of use rather than usage. The problem is that such a specification is almost meaningless. This is particularly evident when one realises that "items throughout will also test comprehension of specific information points" (p.13). Are we much further advanced beyond the point where 'use' is defined merely by 'usage' in context? This dilemma, and the associated failure to provide any idea of what might be meant by an 'understanding of context', is a feature of all efforts to test by relying on reading 'skills' and a so-called communicative approach (cf. the RSA exams for examples of this)

Davies and Widdowson (1974) put forward the distinction between direct reference, inference, supposition and evaluation questions (cf. also Widdowson 1978 - use inference and usage reference). There seems to be here an intuitive grasp of the fact that some order and principle needs to be established, but again no rigorous methodology emerges.

Direct reference questions (or usage reference questions) have much in common with the technologies of Bormuth (1970) and Anderson (1972).

Inference type questions seem to suffer from two defects: in almost all the examples given (this is especially true of Widdowson 1978) inference seems to be synonymous with recognition of cohesion. Thus inference under this scheme amounts to little more than another aspect of 'usage'. It might be argued that the reader has nevertheless to make connections between parts of text of a different quality to the activity involved in direct reference questions. However, the latter also require connections to be made - between question and text rather than between elements in the text - and it is difficult to see how they represent radically different activities.

The second problem is that 'inference' covers an extremely wide range of activities

which are only just now beginning to be recognised: Trabasso (1980) posits four functions of inference, while Johnston (1983) confirms that "the common 'inference' question is no longer a single question type ... Indeed, many literal questions may involve inferencing ,especially at the lexical level." (p.7)

The implications of this seem to be twofold: firstly, as far as the testing of reading comprehension is concerned, it may be entirely proper to emphasise direct reference questions. There is a tendency to dismiss such items as being dull and "mindless' (Widdowson 1978;102), whereas in fact they may be testing much more than we realise and may anyway be the preferred item type in L2 writing. Secondly, research would benefit from a much more rigorous item-writing technology, even for direct reference questions. The concern in interlanguage studies, for example, with acquisitional hierarchies demands a principled *a priori* method of test construction.

4.4.2.5. Systematic item development

Given the need for systematic item development and clearer specifications of content domain, or at least of the relationship between test items and content domain, how might we proceed?

The problem is to avoid loose definitions so that we do not find ourselves in the position acknowledged by Tuinman (1979;40): "every teacher knows it is possible to ask an easy 'higher order' question and a very difficult 'lower order' question." Even Widdowson (1978;102) recognises that usage reference questions "can be made very difficult by accentuating the syntactic and lexical differences between the prompt sentence in the question and the sentence in the passage with which it has to be matched." Indeed it is is well recognised that item difficulty can make even the simplest text challenging or the most complex text easy; what is needed is some systematic way of classifying this difficulty.

The question is closely connected with the problem of generalisability. Johnston (1983;40) reminds us that reading comprehension assessment is merely a more or less systematic sample of reading behaviour which has been taken for the purpose of informing a decision or statement (administrative, diagnostic, selection and classification etc.) and that the level of specificity we require from the test carries certain implications for the constraints which operate on the test, notably in the area of item sampling: "It is now very difficult to argue that a random sample of questions will give us effective information, since it is clear that certain information in the text is more important than other information, and that sorting one out from the other is itself an important comprehension ability which cannot be taken for granted."

Johnston concludes that a systematic method of item development ("preferably rule-governed") would be very useful. If we can develop formalised ways of representing the information which is in the text, as well as the relationships between information segments, and the structural levels of importance, these may be used to select items in a more systematic manner; if we can classify the type of information or relationship which an item is tapping, then we can begin to consider how the reader is handling various intra-text relationships and perhaps generalise within these areas across texts or reading goals. This knowledge will also help us to define the domain to which one can generalise the outcome of a particular assessment device. (see Johnston 1983;43)

4.4.3. Universe-defined domains

4.4.3.1. Definition

Any universe-defined domain consists of a collection of test items which exhausts the questions it is possible to ask within that domain. For this reason, Shoemaker (1975) is able to point to the identity of the instructional program with its associated item universe. For Shoemaker, an item universe is "a collection of stimuli for eliciting responses from examinees." But there is a problem here which centres round the size of the item universe: "Most item universes associated with instructional programs are too large to be dealt with directly." (Shoemaker 1975;131), which leads Shoemaker to propose the idea of an 'item domain' or, synonymously, a 'workable item universe'. An item domain is a clearly definable and enumerable sub-universe of items extracted through expert selection from the larger item universe; the domain is so constructed that, for all practical purposes, achievement as measured by the domain is equivalent to that defined by the universe. (cf. Shoemaker 1975;131)

The idea of a universe-defined domain can be traced to Hively et al. (1968) who, basing their theory on the fact that what 'knowledge' an organism has may be operationally defined as a functional relationship between certain classes of stimuli and classes of response (ib.;276), developed the concept of an 'item form', this being defined as "the rules for generating ... a set of test items." (p.280)

Within the domain of arithmetic achievement tests this produced 'consistent patterns of components of variance' without any statistical item selection, so that Hively et al. were able to conclude (op.cit.;289) that they could place moderate faith in the item forms as categories which represent distinct, homogeneous classes of behaviour and which may thus provide the foundation for detailed diagnosis and

remediation. However, a universe of simple arithmetic items is relatively easy to construct; the problem with language tests is that the domain is much larger.

Hively (1974) suggests that for reading " ... a quite practical domain ... might consist of the front pages from all last year's local newspapers... [Alternatively] the words on the front pages of last year's newspapers may be grouped according to grammatical categories, phonic rules and frequencies of occurrence. Each of these theoretically important characteristics subdivides the larger domain into smaller ones ..." (Hively 1974;8)

4.4.3.2. Problems

Hively's (1974) example of defining domains for reading highlights the problem with item universes: either the level of generality is too great, in which case one is no better off than before, or else the domain becomes too large to be manageable.

Thus Fremer and Anastasio (1969) in developing an item universe for computer-written spelling tests found that even with forty 'error rules' (e.g."Replace *ie* with *ei*") a large number of implausible misspellings were still generated. The position is always that the tension between generality and specificity must be resolved in a spirit of compromise. The example given earlier of German datives (Black and Dockrell 1980), though not consciously constructed within a theory of universe-defined items but rather within a criterion-referenced conception of large-scale domains versus small-scale domains, represents a good example of what can be achieved through this approach.

German dative pronouns are of course limited and knowable; the test constructed was limited according to pedagogic experience: "it was hypothesised that in a situation where they should be using the dative plural, pupils tended to confuse this with the masculine dative singular (error I), the feminine dative singular (error II) or the dative second person plural (error III)" (op.cit.;78) A multiple choice 12-item test was then constructed with errors of each of these three kinds of distractors. For example:

Ich spiele mit den Kinden (a) Ich spiele mit ihm (error I) (b)Ich spiele mit ihr (error II) (c) Ich spiele mit ihnen (correct) (d) Ich spiele mit Ihnen (error III)

So from the 'universe' of German pronouns we concentrate on a sub-universe of dative pronouns; and within this sub-universe we concentrate on the sub-sub-universe represented by four particular dative pronouns. Now while this

provides valuable diagnostic information, "it is clear that tests of this type are relatively complex to create and analyse. As always, the balance between complexity and utility is a crucial one." (Black and Dockrell 1980;81)

4.4.3.3. Facet theory

Related to the concept of item universes is that of structural facet theory, the principal exponent of which is Berk (1978). It should be made clear at the beginning, however, that this approach to item construction is, more than most, applicable mainly when one wishes to test subject matter knowledge. This does not exclude it from consideration but places it at a relatively advanced stage of L2 test construction – when, for example, one wishes to test knowledge acquired on a course of instruction through the medium of the L2 rather than the L2 itself.

The theory posits two structures: content and statistical. The former is concerned with the specification of a research domain that uses a 'semantic structure' known as a "mapping sentence". This then serves as a framework for the statistical structure. According to Berk (1978;62) the mapping sentence is a mechanism for defining a content domain and generating a set of test items to measure achievement in that domain. The mapping sentence is crucial and consists of fixed and variable parts. The fixed part, resembling an item form shell, is partitioned into categories of content or behaviour called 'facets'. These are the dimensions of the domain along which the potential items for measuring knowledge of that content may vary. Each facet is divided into 'facet elements' and is defined in terms of the specific information to be tested. This is presented in the form of a list – the test items are generated by substitution of different combinations of elements in the sentence according to a set of rules. Each substitution pattern produces an item that measures a specific characteristic of the domain.

Vickers' (1973) test for assessing students' knowledge of FORTRAN is a good example of this technique; the method has the advantage of producing valid distractors as well as a vast number of randomly parallel tests from a limited stock of items.

For Vickers, the major disadvantage was that tests generated by this approach required that the student only know how to recognise the various elements of the programming language and not to understand the semantic meaning of the constructions. This need not, however, be a disadvantage in L2 reading tests, since by definition we are not interested in whether students can manipulate items or not. Drill- type exercises lend themselves readily to an approach of this sort, and

traditional substitution tables are examples of facet theory in practice, albeit without a testing theory-driven rationale.

Inadequacies that arise from this approach can be explained if one examines the full implications of facet theory as outlined by Berk. Firstly, objects (by which Berk means all concepts to be learned or behaviours to be demonstrated) should be classified by all properties or facts that the test-maker has chosen as relevant. This immediately raises the problem in language testing that we can never fully isolate elements to be tested; in the German dative test already discussed, for example, one 'contaminating' factor was the gender of the stimulus noun, another was the vocabulary itself, whether known or not.

It is clearly impossible ever to test structure uncontaminated by, at the very least, vocabulary. Therefore, our attempts exhaustively to classify linguistic elements in this way, for teaching and testing purposes, may be doomed to failure.

Secondly, Berk requires that each facet be divided into an exhaustive set of categories or elements.

Thirdly, the elements of each facet should be mutually exclusive; this is one of the hardest requirements of all for language study where, in the nature of things, interaction between elements is a *sine qua non*.

Berk's final requirements relate to the fact that facet elements should be ordered in some way, with the implication that a hierarchy of elements be established. We have already argued that this may not be possible in language on *a priori* grounds (see chapter 2); at the very least, there is no helpful causal hierarchy that can be established among linguistic elements and we shall be driven, as is the case with substitution tables, to limited lists of language items that serve little purpose so far as the generation of test items is concerned. At any rate, little can be achieved beyond the sentence.

4.4.3.4. Testing vocabulary

It would appear that vocabulary is a suitable field for the application of universe-defined tests: one can always produce lists of vocabulary required for learning - Black and Dockrell (1980) give an example of the French vocabulary of farming; what is important in such cases is that an actual list of words is produced, since it is not use merely specifying the subject matter area from which the vocabulary will be drawn. The problem, as always with universe-defined tests, is that

rote learning may be encouraged at the expense of other, more profitable learning strategies.

Sternberg et al. (1983), discussing the teaching and learning of vocabulary within a computer-assisted environment, point out the dangers associated with rote-learning: "in all cases there is heavy reliance upon associative memory ... Poor elaboration in many definitions also reduces opportunities to develop an externally connected cognitive structure." (p.123) On the other hand the learning of word lists, they claim, typically leads to the formation of some degree of subjective organisation, such that recall of a new word or a new word and its definition also touches off recal of other words and their definitions.

Perfetti (1983) cites evidence that context-independent word coding is that which distinguishes skilled from less skilled readers, rather than the use of context.

Our present concern is whether the wash-back effect of testing vocabulary is likely to be counterproductive. It is often recognised that vocabulary test scores correlate highly with general measures of verbal ability, academic success and even with general intelligence. As Perfetti (1983;152) puts it: "a large and precise vocabulary can be considered the public test of the educated person". Now, the danger is that (a) we will rely on vocabulary tests to measure other aspects of language use; (b) students will concentrate on the learning of vocabulary to the excusion of all else; (c) with the result that vocabulary tests no longer measure other aspects of languge use. This is a common fallacy in language tests, to ignore the fact that a selected response test of, say, writing ability is only a valid measure of that particular construct if the ability has been learned in the first place. Otherwise it is merely an aptitude test, if anything. This is the danger of universe-defined tests which rely on vocabulary.

On *a priori* grounds it should be evident that this is so, while empirical evidence comes from Freebody and Anderson (1983), who caution against interpreting high correlations between vocabulary tests and general tests of reading proficiency as indicating that word knowledge is of instrumental importance in text comprehension, mainly because "it takes a surprisingly high proportion of difficult vocabulary to produce reliable decrements in comprehension measures" (p.293). In effect, Freebody and Anderson (1983) have put paid to any interactive reading theory which encourages contextual guessing as a major strategy.

4.4.4. Item transformations

4.4.4.1. Reading comprehension and the pre-eminence of text

In attempting to test reading comprehension it is dangerous to take as axiomatic the need to define comprehension in terms of mental processes. Blanton's (1984) assumption that we should test structures of mind looks as if the L2 reader is being treated as an uneducated half-wit: "Pedagogically the model assumes that second language learners need to acquire the cognitive skills that skilled native readers of English have" (Blanton 1984;37) A 'skills' approach necessarily relies on such a belief since, for example, 'predicting what will come next' is indisputably a mental event rather than something residing in the text.

The problems associated with such an approach may be summarised as follows:

1. Items produced in this way tend to ignore specific language skills;
2. There is a lack of objectivity about such an approach - even such a simple task as identifying the main idea of a passage can produce conflicting opinions (cf. Zuck and Zuck 1984)

In short, we wish to return to a concern with language as manifested in text rather than remain in a world where comprehension is a set of mental processes which operate independently of language.

4.4.4.2. The theory of item transformations

We take 'item transformations' to refer to the process of forming questions to be asked of a text by the act of using some sort of pre-defined *procedure* (an algorithm, perhaps) which is applied to the sentences of the text in a regular manner. The essential ideas behind item transformation technologies in this sense are to be found in Bormuth (1970) and Anderson (1972). Bormuth (1970;39-55) recognises two broad categories of item: sentence- derived items and discourse-derived items. The former subdivide into echo questions (taking a sentence from the text and adding a question mark), tag items (adding a question tag to sentences from the text), Yes/No items, and Wh- items (obtained by deletion of Nouns and Noun Phrases etc.) The latter subdivide into anaphora (pro- words, deleted modifiers, ellipsis, semantic substitute) and intersentence syntax.

The theory of comprehension questions so outlined is elaborated in Bormuth (1969;56) where the analysis posits: rote questions (wh-, tag, yes/no, inflectional),

transform questions (e.g. active to passive), semantic substitute, compound questions, semantically cued questions, anaphoric questions and intersentence relationship questions. Even this relatively crude system represents a considerable advance in the construction of reading comprehension questions; we might compare, for example, the essential looseness of Widdowson's (1978) definitions of usage reference questions: "They direct [the reader's] attention to a particular sentence in the passage and he provides a correct answer by noting how the signification of the sentence of the question relates to the signification of the sentence in the passage ..." (p.100)

Anderson (1972) has the slightly different and less rigid classification of items into:

1. verbatim
2. transformed verbatim
3. paraphrase
4. transformed paraphrase

Anderson points out (1972;168) that we must have firstly a documented rationale for selecting questions to be asked, and secondly a fully explicated analysis of the relationship between the questions and the preceding instruction (or the comprehension question), neither of which we at present possess. As with universe-defined tests of syntax, such needs would require a detailed pedagogic grammar and sound error analysis; thus, we should be able to create tests as measuring devices that form a part of scientific enquiry and are not just a collection of convenient questions.

The intersentence syntax question type has been analysed by Roid and Haladyna (1982;109-110) and leads to a classification based on relations of conjunction, time, cause, concession, illustration, sequence, parenthesis, topic comment, and dialogue.

4.4.4.3. Applications of an item-transformation technology

Most of the practical work relating to item transformations has been concerned with the analysis of instructional texts i.e. texts in which information is being imparted rather than texts designed to test specific language skills. As such, the implications for L2 testing are not as clear as they might be.

Thus, Roid and Haladyna (1982) give the example of transforming a statement of principle into a question: Premack's Hypothesis that "Given any pair of responses the

more probable one will reinforce the less probable one" becomes transformed, eventually, into the test item: "When there are two activities, stacking two cords of wood and watching a championship football game, which activity will increase the chances that the other will be repeated in the future?" (p.112) In practice, this appears to differ little from the item-forms approach of Hively et al. (1968).

On the other hand, Roid and Haladyna (1978) in a study designed to examine the effectiveness of using rules for item-writing which were similar to but not as rigorously automated as those suggested by Bormuth (1970), found that neither the writing of items from instructional objectives nor the use of rules for transforming instructional sentences into test items completely removed subjectivity from item writing, which is not really surprising.

More importantly, from a theoretical point of view, they found that rule- generated items were easier to answer than objectives-based items, regardless of the item writer or the test occasion. It should be emphasised that this was with 'instructional' material and no comment was made on the effectiveness of questions so far as language-specific comprehension was concerned.

Roid et al. (1978) showed that the method of selecting the 'question word' (i.e. which noun or adjective in a sentence was to be deleted or transformed or produced as the correct answer) played a crucial role in determining the pattern of pre-testand post-test item difficulties of the resulting items. The use of nouns that occur frequently in a passage was shown to create items that are too easy if the sentence in which they first occure is used. Rare 'singleton' nouns, which are relatively rare in American English and appear only once in the passage, were found to be the most effective question words in a study.

Finn (1975) applies the notion of an item-writing technology based on transformations of textual elements to the specific area of reading achievement tests. His starting point is Bormuth's proposal that questions be generated from written texts by analysing the syntax of the text and generating questions by rules which refer to the syntax, but unlike Bormuth he rejects an analysis based on transformational grammar because of the "problems inherent in surface structure analysis" (p.343).

Finn sees a solution in underlying structure analysis, by which he means a Fillmore-type case grammar. Now questions can be generated by deleting cases and verbs in underlying structure. Two advantages are perceived for this approach: firstly, expletives, 'functional' verbs, articles and prepositions either appear as part of phrases or they do not appear at all. Secondly, the number of questions it is possible to write

over a given sentence becomes independent of the number of words in the sentence and is a function of the number of case phrases and non-zero verbs in the underlying sentence.

Interestingly, Finn found that questions derived by deleting the verb and the objective are highly anomolous (e.g. the sentence "Nixon forced Hickel to resign" produces questions such as : "What did Nixon do to Hickel to resign?") and are very poor candidates for a test of reading achievement. After testing his 82-rule algorithm with five test item writers, Finn concluded that there is an inherent bias towards Noun Phrases (e.g. there is no easy way to write a question from the sentence "The lamp costs a dollar" to which the answer is "costs"). This can be viewed either as a fundamental limitation on reading comprehension questions, at least so far as extended text is concerned; or it can be seen as a partial definition of reading comprehension: "A subject's ability to recall deleted NP's in answer to grammatically well-formed questions appears to be a legitimate concept of reading achievement." (Finn 1975;359)

This also helps to remove some of the argument over use/usage questions, i.e. in testing use we cannot easily test for verb knowledge, and it must be assumed. Put another way, a question which aims to test specific verb knowledge is likely to emerge as a 'usage' item and be perceived as unnatural, uncommunicative, or whatever.

Finn's algorithmic approach raises other questions about the nature of language; for example the difference between 'mountain tops' and 'navigation laws'; the latter easily generates the question "What kind of laws ...?", whereas the former hardly supports the question "What kind of tops ...?" From this it can be concluded that a rigorous approach to item writing is going to involve investigation of specific uses of language.

4.4.4.4. Criticisms of item-transformation technologies

One immediate problem becomes apparent from Finn (1975): the sheer number of rules required, even for an 'idealised' and simplified algorithm is immense. Finn needed 82 rules in the end, though he started with only 11, and it is doubtful whether time will be invested on this scale for day-to-day tasks.

Roid and Haladyna (1982;93) have pointed out that one of the criticisms of prose transformations is that they can lead to the generation of trivial items: "It is feared that items will be keyed to relatively unimportant ideas or to the verbatim-recall level

exclusively." This is more likely to be a problem when testing on instructional material, since it is by no means certain that straightforward verbatim (direct reference) questions are not of value in testing L2 reading. Indeed, Roid and Haladyna's comments suggest that much of value may be done within L2 testing: "The state of the art of item transformations for assessing reading comprehension is still relatively primitive. Hence, item transformation is open to the criticism that it does little more than test at the comprehension level. On the other hand ... basic terminology and concepts must be understood before higher level thinking is possible in an academic discipline. Therefore it seems that basic exercises in comprehension will always be needed ..." (Roid and Haladyna 1982;97-98). One might add that the legitimate concern of L2 testing is only "the comprehension level".

Cronbach (1970) applauds the notion of universe-defined tests and item transformations but adds "... this does not help us much in thinking about universes carrying more interesting content." (p.510) This again depends on the view that item transformations cannot be used profitably at the lower level of 'comprehension'. Diederich (1970;1005) also focuses on the use of item transformations for instructional text and, legitimately, raises the problem that we do not know theoretically how many items would be allowed on one sentence, leading to the situation where we could get at least 960,000 items from one short physics book. Shoemaker (1975;134) similarly suggests that we do not know how to determine therelative importance of each item.

In part these criticisms are answered by pointing to algorithmic techniques as developed by Finn (1975), where an effort is made to establish reasonable grounds for choice between possible items.

Lucas and McConkie (1980) argue that an approach to test item writing based on Bormuth's (1970) proposals is limited by the ability to "deterministically assign structural descriptions to texts" and that in addition "the approach inherently provides only limited information about the questions described." Specifically, indicating the part of a text from which a question was derived is not the same as specifying which parts of the text are pertinent to answering it, partly because of the well-known redundancy of prose and because of a reader's ability to draw inferences about unstated information. In many cases, argue Lucas and McConkie, even deleting from a passage the specific text segment from which a question was generated might have little or no effect on the probability of answering the question.

All of this raises the question of how text might be analysed, to which we now turn.

4.4.4.5. Generalisability

The investigation of item universes inevitably raises the question of generalisability in acute form. Osburn (1968;95) states the matter in its clearest form: "We almost always have some larger universe of content in mind [than the items actually administered on a test], and our implicit objective is to generalise over the entire content domain ... the fundamental objective of achievement testing is generalisation."

As the situation stands at present, achievement tests, at worst, consist of arbitrary collections of items thrown together in a haphazard manner. At best, such tests consist of items judged by subject matter specialists to be relevant to and representative of some incompletely defined universe of content. In neither case can it be said that there is an unambiguous basis for generalisation, for the reason that the method of generating items cannot be stated in operational terms. The temptation to resort of statistical strategies is popular because it invokes "the concept of a latent variable - an underlying continuum which represents a hypothetical dimension of knowledge of skill." (Osburn; loc.cit.)

Thus by taking a collection of items and referring them to a latent variable with a name, we create the illusion of generalisation. Osburn reminds us that the basis of generalisation must be contained in the operational definition of the procedures used in generating and sampling items that go to make up the test, so that in a universe-defined test "one simple way to specify a finite universe of test items is to make a catalogue of all items that you wil allow to appear on the test. For example, a word list of the 5000 most frequently used words might constitute an explicitly defined finite universe of content for a spelling test." (op.cit.;97)

It is apparent that one of the problems with universe-defined tests in language is that testees may be tempted, if enough is at stake, to learn the individual items by rote. This has occurred in military schools using the American Language Course Placement Test, where the domain of the test is explicitly defined by the vocabulary and grammar of the course book. This may be viewed as the generalisation question in reverse - testees fail to learn properly because they do not trust their own powers of generalisation from knowledge to test performance, and because the test items can be answered by applying a rote learning strategy. What has in fact happened here is that the domain of L2 reading has become, unintentionally, isomorphic with the domain of L2 vocabulary; an example of misguided test construction.

4.4.5. The analysis of text

4.4.5.1. Descriptors

An alternative approach to a Bormuth-type analysis of text is proposed by Lucas and McConkie (1980;134), which requires that the passage to which questions are to be related be segmented into units of sufficient detail for the user's needs, with each unit numbered for referential purposes; content is represented using propositional network structures (cf. Frederiksen 1975). A set of ten descriptors then relates questions to passage content units (which are not meant to be exhaustive or definitive).

To describe a question using this system one first identifies each proposition judged to pertain in any way to answering the question. The criterion to be applied here is that a proposition should be included in the description if the information represented by the proposition could in any way be used by a reader to answer the question being described. One or more of the ten descriptors is then used to describe the manner in which the proposition relates to the question. The set of all such proposition/ descriptor pairs constitutes the description of the relation of the question to the passage.

A complete account of the descriptors is given in Lucas and McConkie (1980; 135-139); the four 'primary' descriptors are, for example, Stated [information], Implied, Inferred and Assumed. The value of such an approach lies in the meaning it gives to the term 'difficulty level'. The complexity of a question now resides not just in the question itself or in the passage but in the relation between the two. This intuitively satisfying idea is thus given firm theoretical backing.

4.4.5.2. Identifying high information words

An emphasis on developing questions that measure important aspects of a prose passage (i.e. use rather than usage, value rather than signification) requires an objective analysis of text. Roid (1979) cites certain follow-up studies by Finn based entirely on word counts (e.g. the word 'the' occurs once every 10 words, 'incarnation' once every billion).

High information words tend to be those which are relatively rare in the language as a whole and which occur only once in the passage; they tend to be words which are difficult for students to guess if they are deleted from a prose passage as in, for example, a cloze test. Indeed, standard frequency and text frequency have been

shown by Finn(1977) to predict the ease with which a deleted cloze item may be restored. Verbs and adverbs, although they may be high information words, are not good candidates for question words, whereas Adjectives, Nouns, Adjective Phrases and noun Phrases are.

Roid's own experiments and her review of others led her to the conclusion (1979;86) that the concept of identifying high information words from prose passages for use in selecting sentences to be transformed into questions is a workable methodology. Words identified as rare singletons, which have a standard frequency index of 60 or less and which occur only once in a prose passage, are useful candidates for questions that test learning from prose. Such ideas need to be applied more widely in L2 reading tests before we can comment with authority on their worth in our own field of interest.

4.4.5.3. Topic and the structure of text

The notion of topic is of central importance in the analysis of text, especially when one begins to think in terms of automated question writing. We have already seen the problems which can arise when different groups try to identify the 'main idea' of a passage (Zuck and Zuck 1984). Item transformations have not yet progressed sufficiently to provide an operational definition of topic, and indeed may never do so, since, in the nature of things, a topic does not necessarily relate to the words on the page but may be the *a priori* organising principle of a text (see, for example, Bransford and Johnson 1973 or Anderson et al. 1977).

The need to consider 'topic' arises from the fact that many approaches to text analysis, and particularly Bormuth-type item transformations, depend on a perceived hierarchical organisation of discourse and thus imply that 'topic' may be represented by the top-most elements in the hierarchy (cf. Bormuth 1970;55). Unfortunately there are serious objections to this view on theoretical grounds, stemming largely from the fact that despite the appearance of a highly formal and therefore objective type of approach, "the proposition-based analysis of natural language texts is inevitably subjective" (Brown and Yule 1983;114). Brown and Yule conclude that formal attempts to identify topics are doomed to failure (op.cit.;68).

The reasons for this are clearly set out in Jackson (1984), where three major ideas are discussed:

1. a topic is considered as a system of concepts; "what distinguishes a topic from a random collection of related ideas is

goal-directedness."(p.22)

2. an expository text is viewed as a linear formalisation of a topic; because it is essentially linear the text therefore represents a linear formalisation of a topic, insofar as it seeks to describe the conceptual system in propositional terms, and cannot proceed without introducing such notions as beginning and end, earlier and later, even though these may have no obvious counterpart in the structure of the system. The formalisation is thus distinct from the system it describes. (p.23)

3. The structure of a text is seen as a pragmatic realisation of the structure of its topic: "the function of expository text is not merely to state truth or present facts; it seeks to communicate and promote comprehension. The order in which statements are made and questions are posed is therefore dependent upon more than the relationship between signs and the world (semantics). At least as important is the relationship between signs and interpreters." (p.23)

One might argue that this latter point reintroduces the problem of how much weight we give to the reader and how much to the text, which we have already decided to resolve in favour of the text. However, the results of present work on investigations into topic suggest that dependence upon higher elements in a 'discourse tree' may be mistaken and that complete technologies of item-writing for units beyond the sentence may be a fruitless pursuit.

4.4.5.4. Identifying the structure of text for problem-solving

Assuming that a reader can 'read' in the sense that he can assign meaning to symbols on a page, up to individual phrase and sentence level, but is unable to 'read' in that he cannot combine these meanings in a logically consistent way and so fails to draw inferences from written material, fails to detect inconsistencies in various parts of a literary message, or fails to determine whether additional new material is logically independent of what has already been read, how might we set about testing in these circumstances?

We should proceed with caution, since it is by no means clear that problems in L2 reading arise for such reasons; the cumulative effect of failure to identify word meaning and sentence structure may lead us, falsely, to assume that a failure of inferencing is at issue. Nevertheless, we need some means of classifying text tasks from this viewpoint, especially if we take Moffett's (1968) view that critical reading is actually critical thinking about written discourse.

The clearest attempt at constructing a framework from this point of view is seen

in Scandura (1977). Relying on the fact that all lists of skills involved in critical thinking (or critical thinking while reading) include skills involving the ability to reason deductively and that the number of basic logical inference rules is relatively small (no more than 24 are sufficient for all proofs in first-order logic) Scandura identifies various dimensions over which reading materials may vary. These were determined by 'strictly analytic means' and there was no use of factor analysis or other statistical procedures. Dimensions within a level were assumed to be hierarchical, but there are no assumptions made about the relationships between levels.

The resulting outline looks as follows (from Scandura 1977;413):

Figure 2
Dimensions and levels over which reading materials may vary

1. Level A. Relation of statements in message to reality.

    a. Statements that agree with facts known by the reader

    b. Statements that neither agree with nor contradict facts known by the reader (neutral statements)

    c. Statements that contradict facts known by the reader

Level B. Complexity of context (including length)

    a. Simple; single implication; message containsonly relevant statements

    b. More than one implication; message contains only relevant statements

    c. More than one implication; message contains 3-5 extra statements, 1 or 2 of which may appear to be relevant.

Level C. Availability of premises in message

    a. All relevant premises present and clearly stated Nuance; premise determined from context

    b. Some premise is missing but implied by the context

Level D. Required length of chain of inference

    a. Single rule application

b. 2 rule applications

c. 3 rule applications

d. 4 rule applications

Level E. Terminology used

a. Most common English terminology (e.g. "If A, then B")

b. Variations from common terminology (e.g."Only A if B", or "B is necessary for A")

In essence, then, the difficulty levels depend on the amount of information that must be processed and how consistent the information is with what is commonly known. Scandura reports that an attempt was made to identify dimensions that are independent for purposes of classifying reading materials, but that "no definitive claim can be made regarding their behavioural independence' (p.415).

However, as a result of the small-scale experiment Scandura was able to conclude that "... dimensional analysis appeared to be a useful technology. The testing assumptions were shown to be valid, and the tests were shown to be efficient instruments for measuring the reading contexts in which children can use a logical rule." (p. 424) Such an analysis could provide a useful measure of difficulty for L2 reading texts.

4.5. Conclusion

Text and task are clearly inextricable, and it makes sense for us to talk of 'text and task' difficulty rather than either 'text' or 'task' difficulty independently of each other. Technologies of test construction are useful aids in the process of systematic item construction and have much to offer in the realm of criterion-referenced assessment. Indeed, in many cases the two pursuits (criterion-referenced assessment and developing item-writing technologies) differ hardly at all. Their importance in the item-banking context is that such methods enable us to be as rigorous about the 'dimensions' of our bank as we can be at the pre-test phase. Another way of putting that would be to say that such methods go as far as is possible towards ensuring content validity at the test writing stage.

'Technologies' of the type discussed here are of course open to criticism, particularly in so far as they seem to fragment the subject area to an unwarranted extent and also tend to suggest that learning is a simple incremental process.

Nevertheless, used in the appropriate manner and in the appropriate context, such technologies can be a powerful means of ensuring content validity in testing for many different purposes.

CHAPTER 5

ITEM BANKS: PRACTICE AND PROBLEMS

5.1. The statistical background

5.1.1. Introduction

Most of the problems associated with the development of item banking have had less to do with the content considerations we have been looking at in previous chapters, important though these considerations are, and more to do with technical matters, in particular the problem of which statistical model to use and how to calculate the parameters of that model.

Such technical considerations have tended to obscure the discussion of test content. One of the main themes of this thesis is that content considerations are just as important in item bank development as in in more traditional test development. Indeed, content assumes a greater importance in item banking because of the whole issue of dimensionality. We leave this issue now, however, to look at the technical problems which face any item bank designer.

5.1.2. Which model?

Essentially, we have to choose between three models, depending on the number of parameters we wish to include in our analysis. There may be variations in the exact implementation of these models (for example in the exponential factor chosen: most models use the natural logarithm base e as their exponent, but Choppin (1978) for example uses $W-$ an exponent designed to produce a more meaningful scale), but in the last analysis we are looking at only three models, all of which attempt to model an examinee's performance on a test item as a function of characteristics of the item and the examinee's ability on some unobserved, or latent, trait. The IRT model specifies the relationship between a latent trait and observed performance on the test that is designed to measure that trait. This relationship in mathematical form is usually referred to as an item characteristic curve (ICC), or an item response function (IRF).

The first type of model is the one-parameter logistic model; this is more usually referred to as the Rasch model, after its developer (see Rasch 1960 for a full account of the development of this model). It is the simplest model and the most popular. It may be expressed as follows:

$$P_g(\theta) = [1 + \exp(-(\theta - b_g))]^{-1}$$

where P_g represents the probability of a correct response to item g by an examinee with ability θ and b_g is the difficulty of item g. This formula does no more than express the relationship between ability and difficulty as a function which is more likely to be near 1 when the examinee's ability is much greater than the difficulty of the item, and more likely to be near 0 when the examinee's ability is much lower than the difficulty of the item. When the ability of the examinee matches the difficulty of the item the value of the function will be 0.5.

The second type of model is the two-parameter logistic model, first developed by Birnbaum (1968). This has the following form:

$$P_g(\theta) = [1 + \exp(-Da_g(\theta - b_g))]^{-1}$$

where a_g is the discrimination parameter for item g, D is a constant equal to 1.7 and the other parameters are the same as described above. This model takes into account the observation (or hope?) that not all items will discriminate equally and that we should expect ICCs to cross and not to have identical non-overlapping slopes.

The final type of model is the three-parameter logistic model, most closely associated with Birnbaum (1968) and Lord (*passim*). It has the following form:

$$P_g(\theta) = c_g + (1 - c_g)/[1+ \exp(-Da_g(\theta - b_g))]^{-1}$$

The only difference between this model and the two-parameter model is that it includes the parameter c_g which describes the lower asymptote of the function (or if you like, the lowest y co-ordinate) and is best thought of as a chance scoring, or guessing, parameter. For example, in a multiple choice test with four options, the c_g parameter would allow us to take into account the fact that an examinee could get a 25% result purely by chance. This is something which the other models, and particularly the Rasch model, cannot allow for - the assumption must always be that there is no chance or guessing element involved.

In all three models, difficulty and ability parameters are expressed in log odds units called logits. These are usually centred on zero and have a standard deviation of 1. Thus an ability logit of -2.3 shows a very low ability examinee, a difficulty logit of -2.3 shows a very easy item. Both ability and difficulty will be expressed on the same scale.

Which model should be used, then? Proponents of the three-parameter model

argue that chance scoring is a reality of multiple choice items and should be used when a test of this type is being analysed. On the other hand it has to be said that it is impossible to estimate the c_g parameter accurately and anyway guessing is a characteristic of the examinee not the item. Moreover, while the three-parameter model often produces the 'best' estimates of parameters, this does depend to a large extent on the distribution of ability in the sample being studied (Ree 1979): if ability is uniformly or normally distributed then parameter estimation can be relied upon, but skewed ability produces less reliable results.

More importantly, perhaps, the size of the sample has an important bearing on the effectiveness of parameter estimation. Reckase (1979) has shown that in using Rasch and three-parameter methods of estimation for calibration of an anchor test (see later in this chapter) then stable linking occurred with a sample of 300 for the Rasch model, but a sample of at least 1,000 was needed for the three-parameter model. With small samples the three-parameter model (at least in the LOGIST implementation of it) yielded extreme difficulty values and discrimination values near zero. Hulin *et al.* (1982) have also shown that there is a trade-off between test length and sample size: for estimating parameters from a three-parameter model a data matrix of 60 items by 1,000 examinees produced stable results. By halving test length and doubling sample size comparable precision was achieved.

One of the problems in using the three-parameter model, then, is that a large amount of data is required in order to achieve stable parameter estimates. In addition, examinee abilities need to be spread sufficiently normally across the range of difficulty of the test. Choppin (1978;15) suggests that the three-parameter model should not be used because the additional complexity of analysis required to establish the estimates is counter-productive. While working with the Rasch model, he claims, it is sufficient to obtain data from "a few hundred people" to arrive at a fairly precise estimate of an item's behaviour.

The other factor influencing our choice of model will be our opinion as to the importance of the 'discrimination' parameter. Choppin (1978; ibid) suggests that conceptually for any two items *i* and *j* one "should be consistently harder or easier than the other across the whole ability range" if they are to form part of an objective measurement system. The differential discrimination implied by the presence of the discrimination parameter again seems at variance with the underlying idea of much IRT theory, namely that relative difficulties remain constant. This latter view is forced upon us if we adopt the Rasch model (hence much criticism of this stance), but the argument from complexity will still apply, and we are still faced with the conceptual

problem that IRT tolerates a discrimination parameter with difficulty. The argument is essentially as follows.

Item characteristic curves for unidimensional tests cannot intersect (which is what the discrimination parameter implies) since this would imply that two items will have different orderings of difficulty for subjects of different ability (Lumsden 1978; 22). It is possible that with two items both measuring height one is harder for shorter subjects and the other harder for tall subjects. Therefore test scaling models are self-contradictory if they assert both unidimensionality and different slopes for the ICCs. The two- and three- parameter models should therefore be abandoned since if the unidimensional requirement is met, the Rasch one-parameter model will be realised.

The problem with *this* view is that there is only slight empirical evidence for the assertion that ICCs for the items of a unidimensional test will all have the same slope (Lumsden op.cit.; 24)

Where does this leave us? Much work has been done with the Rasch model to see whether parameter estimation is compromised by using only a two-parameter model (see for example Wright and Stone (1979), Willmott and Fowles (1974), Skaggs and Lissitz (1986)) and whether it works in practical situations (Baker (1987)). Given the relative robustness of Rasch parameter estimate procedures, the difficulties outlined above of using the three-parameter model, and the fact that the Rasch model is the simplest to use, we prefer to use the two-parameter Rasch model. This does not mean of course that we do not recognize the difficulties involved with using this model, but provided our assumptions are made clear we feel that progress can be made.

5.1.3. Ability and difficulty: parameter estimation

The parameters of 'ability' and 'difficulty' (as well as 'discrimination' and 'guessing' in the more complex models) can never be determined empirically, they can only be estimated (Rasch 1960; 77; Hulin *et al.* 1983; 99). Moreover, there is not a mathematical proof that estimates of item and ability parameters obtained simultaneously become more accurate as test length and sample size increase, though estimates do improve provided that there is a reasonable match between item difficulties and examinees' abilities. In fact, the best estimate of the ability parameter can be derived from the examinee's raw score, and the best estimate of the difficulty parameter can be ~~be~~ derived from an item's score (i.e. proportion correct, or facility value), which leads directly to such observations as "the higher the number of

candidates giving the correct answer to an item, the smaller is its estimated difficulty, so that the rank order of items by estimated difficulty is the same whether the estimation is carried out by assuming the Rasch model or by calculating the simple proportion of correct responses to the item" (Woods and Baker 1985; 126). This also means, in the terminology originally employed by Rasch (1960), the candidate's raw score is a sufficient statistic for estimating the ability parameter, while the item score is a sufficient statistic for estimating the difficulty parameter.

In practice, then, what the estimation of ability and difficulty parameters involves is essentially the normalisation of skewed distributions so that, for example, a set of items calibrated on a high ability group will be 'normalised' so that the difficulty estimates are adjusted downwards. It is quite clear that this is happening in the formulae given by Wright and Stone (1979; 18ff), which form the basis of the widely used BICAL program.

If this is the case, what is the difference between Rasch estimates of parameters and the traditionally reported test statistics? So far as the raw score/person ability value is concerned, one important difference is that person ability is reported on the same scale as the difficulty values, so that not only is a comparison possible, but also a matching process can take place. So far as the facility/difficulty value is concerned, there is indeed a one-to-one relationship between the Rasch item difficulty value and the item p-value, but the Rasch value includes a transformation so that the p-value, which is not linear in the implied variable, becomes linear. Another important difference is that the standard error of measurement of the p-value is greatest in the middle of its range (i.e. at p=.5) and zero at the extremes of its range; the standard error of measurement of the Rasch difficulty estimate on the other hand is smallest at p=.5 but goes to infinity at the extremes. This reflects what is really common sense - that we can not gain any information from a perfect or from a zero score (a perfect score does not mean, for instance, that the examinee knows 'everything' any more than a zero score means that he knows 'nothing').

5.1.4. Invariance across populations

As we have seen, one of the much-vaunted claims made for the Rasch model in particular and for IRT in general is the fact that it possesses the property of 'invariance' - invariance across items and invariance across populations. Estimation of parameters is said to be 'sample free' for estimates of ability and 'person free' for estimates of difficulty (see Wright and Stone 1979 for example).

The first important point to make is that this property of sample-freeness derives

directly from a simple arithmetical property of the formulae which form the basis of the model. Once we remove the labels 'ability' and 'difficulty' from the parameters of the equations and treat them simply as algebraic entities, then we can see quite clearly that by expressing one entity in terms of the other (and the two terms should always be defined simultaneously - see Rasch 1960; 73) then what we have is in fact a pair of simultaneous equations which can be solved in the usual way by eliminating one of the unknown expressions and including it in terms of the other. There is nothing mysterious about this (the arithmetic will not be shown here, but it can be found in Choppin 1978). 'Sample-freeness' is thus, in one sense, another aspect of the attempt to provide a 'semantic' meaning for a 'syntactic' expression (Lord and Novick 1968), and we should not be deceived into thinking that it is some sort of wonder ingredient.

The second point to make is that the whole idea of 'invariance' seems counter-intuitive (indeed this notion is explored at length in criticisms made by, for example, Goldstein - see later in this chapter): it seems sensible to ask (Skaggs and Lissitz 1985; 519) how different samples can be before they should be considered as separate populations. In operational terms, this is a question of how different IRT calibrations will be between different groups of examinees or similar groups of examinees at different times.

Woods and Baker (1985) and Baker (1987) present empirical evidence which seems to show that the estimation of the difficulty parameter is relatively stable across what they claim to be very different populations, namely groups of Tanzanian learners of English (a low ability group) and Malaysian learners of English (a high ability group). What these studies show, however, is not so much the invariance of the difficulty parameter estimate as the variance of the classical facility value. This latter, however, was never intended to be interpreted in a fixed way, and it is misleading to claim invariance for the (Rasch) difficulty estimate by comparing it with the classical facility value.

Other studies have been less conclusive. Harris and Kolen (1985) compared equating results based on high ability examinees with those based on low ability examinees for five forms of a mathematics usage test. In all cases a negative bias occurred, that is to say the low ability samples produced higher equivalent scores than the high ability sample. Cook *et al.* (1984) investigated the stability of the three-parameter model calibration at two different points in time. The results showed that item difficulties were more similar between the autumn old form and autumn new form calibration than between the autumn old form and the spring new form

calibration (of a biology test where systematic differences were known to exist between the autumn and the spring samples in terms of recency of instruction in biology). The conclusion was that the properties of an achievement test can depend on the calibration sample. In particular, recency of instruction seemed to be a key variable differentiating samples.

These findings are among the most interesting in relation to the IRT claim for invariance and sample-freeness since they provide empirical support for the view that at some point samples must be treated as separate populations. Skaggs and Lissitz (1986; 520) go so far as to suggest that calibrations based on samples in different parts of the country are probably not comparable. The implications are far-reaching: anyone using and IRT-calibrated item bank would be well advised to recalibrate items based on local samples if scores derived from ability estimates are to be used. Likewise items should be recalibrated regularly.

This weak conclusion is in fact the position which now tends to be adopted - Woods and Baker (1985) and Theunissen (1987) both tend to play down the invariance property of the Rasch model. The problem with this is that we are left with the serious question of whether it is worth all the effort in implementing IRT models if we cannot make much use of one of its major claimed properties.

5.1.5. Dimensionality

Actual tests are never perfectly unidimensional in that different parts of the test, and different items within the test, are usually designed to test different aspects of a topic. To a certain extent, then, the problem of unidimensionality is a problem of definition - a test may be said to be unidimensional if we *say* it is unidimensional, this being a matter for content validation studies. From one perspective, the dimensions of a test emerge as we go closer into the test content; for example, 'language proficiency' may be our single dimension (when compared with, say, mathematics or music), in which case we would not seek further dimensions. Or we may choose to divide this construct up into a number of other dimensions which reflect current thinking on the structure of language proficiency - into the dimensions of productive and receptive skills for example. These may be further subdivided (into reading and writing, say), and we may choose to go further still and say that 'reading', say, is a dimension on its own, quite separate from writing. But a reading test of any kind presumably is composed of test items which are designed to test different aspects of what reading involves, so are we justified in thinking of reading as multidimensional?

In practice, problems have arisen in areas such as maths, which can be thought of as a single though diffuse area of study (Choppin 1978; 19–21). A general measure of achievement may be thought to be adequate, so the simple way of dealing with the situation is to create a set of separate sub-systems within the overall measurement system. Each latent trait is represented by a collection of perhaps one or two hundred test items which are cross-calibrated one against the other. The result will be a profile of performance.

But by scaling subsets onto a common latent trait one is ignoring aspects of the subtrait which are unique to it. Thus 'geometry' when analysed as a part of general mathematics measures that geometry performance which is similar in some way to other 'mathematical' behaviour. The unique abilities required to solve geometric problems can only be given due weight when 'geometry' performance is scaled as a separate sub-trait. Similar arguments would apply to the components of language/reading proficiency.

One thing is clear, however, and that is that it is never clear what degree of multidimensionality in the item responses could be tolerated before IRT methods should be ruled out. On the other hand, it is a requirement of all IRT models that they be used on unidimensional tests. So how should we approach the problem?

The simplest approach is simply to define a unidimensional test as a test in which all the items are measuring the same thing (Lumsden 1961; 122). This, however, seems unsatisfactory in so far as it begs the question. A next step on from this is to consider the answer pattern that would be generated by unidimensional test with infallible items. If the items are arranged in order of difficulty, then a kind of implicational scale will be seen, in that an examinee passing the first item and failing the second will fail all the rest, an examinee passing the first n items and failing the $n+1^{th}$ will fail the rest. This holds up to a certain extent, but will not do if there are in fact two or more dimensions for which a strict hierarchy of difficulty holds i.e. all the items on dimension 1 are easier than all the items on dimension 2 and so on.

A more complex definition of unidimensionality (or strictly speaking of dimensionality of any order) is found in Lord and Novick 1968 (p.359):

> Consider a set of k items and one latent trait ξ which affects examinee performance on all items in the set. We can now represent each examinee as a point on the trait. Next consider all the examinee populations that may be of interest for this set of k items. Assume that each item is administered just once to each examinee, and consider the conditional frequency distribution (over people) of item score for any

> fixed value of ξ. If this (unobservable) distribution is not the same for all the populations of examinees, then there must be one or more psychological dimensions in addition to ξ that discriminate among the populations of interest. In defining the *complete latent space* therefore we must include these additional dimensions. *Thus, by definition, in the complete latent space the conditional distribution of item score for fixed x is the same for all populations of interest.*

From this definition of unidimensionality it folows that the ICC for an item is invariant for those populations used to define the latent space.

Another definition (Gustafsson 1980; 207) is more directly related to the Rasch model and to the concept of local (or conditional) statistical independence. In this definition (which is complex mathematically and will not be given here) the probability of an examinee response pattern (which is essentially what determines dimensionality) is given by the product of the probabilities of the item responses, and an individual's performance depends on a single underlying trait if, given his value on that trait, nothing further can be learned from him that can contribute to an explanation of his performance. The proposition is that the latent trait is the only important factor and, once a person'svalue on the trait is determined, the behaviour is random, in the sense of statistical independence. In other words, the Gustaffson definition builds on Lord and Novick by including specifically the idea of local statistical independence.

McDonald (1981) has pointed out that the notion of a unidimensional set of binary items does not possess a widely accepted definition, and that definition must precede verification. What we find (McDonald 1981; 100- 101) are notions of *unidimensional, homogeneous* or *internally consistent* sets of test items and we see that these notions seem to have something to do with certain aspects of item-stem content, or with certain statistical analyses of examinee responses to the items as stimuli, based on various psychometric theories. Yet there is no general agreement as to what the verbal labels signify (which does not necessarily require us to adopt the logical positivist conceptions of meaning as verification and of explication as the arbitrary process of operational definition). If there is a prevailing conceptualisation of the notion that a set of n tests, yielding quantitative scores, is 1-, 2-, or r-dimensional, it is that 1,2, or r common factors explain the correlations of scores from distinct tests in the set, computed over a defined population of examinees.

This is an important observation: in principle, a set of n tests or of n binary items is unidimensional if and only if the set fits a (generally non-linear) common factor model with just one common factor. This does not mean, however, that 'unidimensionality' is a synonym for 'homogeneity'. As McDonald points out (ibid.)

there is no logically necessary connexion between the mathematical conception of a set of unidimensional variables and the substantive conception of a set of tests or items that measure in common just one property of the examinees in a given population. The empirical heuristic of common factor analysis connects the model to its applications in the real world by interpreting the common factor as a characteristic of the examinees that the tests measure in common and a residual as a specific characteristic not measured by any other test, as well as, possibly, an error-of-measurement. The same heuristic is employed in applications of latent trait theory, except that the item specific is not usually given an explicit definition.

However, the fact that factor analysis is so closely connected with the investigation of dimensionality of the latent space means that, at least in language testing terms, we shall be drawn into discussions of this substantive kind, since this is how we have chosen to investigate the structure of language proficiency (at least in part).

Most investigators now recommend some form of factor analysis as the cornerstone of any procedure to assess dimensionality. Hulin *et al.* (1983) and Hambleton and Swaminathan (1985) are among the more recent to advocate such methods. What criteria should we adopt to accept the unidimensionality of a test? Two simple methods emerge, which will be outlined now, though it should be noted that they are in essence two different ways of doing the same thing.

First, there is the so-called Scree Test described first by Cattell (1965). This requires us to examine the graph of eigenvalues and stop factoring at the point where the eigenvalues (or characteristic roots) begin to level off forming a straight line in almost horizontal slope. Beyond this point Cattell describes the smooth slope as "factorial litter or scree". In assessing dimensionality the implication is that if there is more than one factor in evidence before the slope levels off (whatever the actual value of the eigenvalue, though usually this will be around 1 or below) then we do not have a unidimensional test. This "root staring" criterion is often criticised for being subjective because, for example, it is not uncommon to find more than one major break in the root-graph and because there is no unambiguous rule to use.

There are two answers to this objection. First, as Kim and Mueller (1978; 45) point out, given the complexities as well as uncertainties inherent in the method, the final judgment has to rest on the reasonableness of the solution on the basis of current standards of scholarship in one's own field. "This criterion is elusive but, fortunately or unfortunately, all of us must live with it in order to communicate our findings to

our fellow scientists" (loc. cit.).

The second answer brings us to the second method of determining the dimensionality of the test we are investigating. Reckase (1979) has provided a more objective evaluation of the eigenvalues obtained so that we can feel more confident about extracting (or not extracting) meaningful factors. Reckase bases his evaluation on a consideratin of the proportion of variance associated with the first eigenvalue and the ratio of the first to the second eigenvalue of the interim correlation matrix. For acceptable calibration, the first factor should account for at least 20% of the test variance; if a factor other than the first factor is of interest, factor pure subtests should be formed and calibrated separately.

Henning *et al.* (1985) have investigated the question of unidimensionality for language tests. Using results obtained from the English as a Second Language Placement Examination, they claim that the 6 subtests of the ESLPE together form a unidimensional test, in spite of the fact that the subtests appear to be measuring separate aspects of proficiency (listening, reading, grammar, vocabulary, writing (error detection) and composition). No attempt was made to allow for student background, indeed the authors conclude that their study shows that the Rasch model may be used for the development and analysis of language tests which may comprise item domains representing diverse subskills of language use and which may be applied in the testing of persons from diverse national, linguistic, cultural, educational and professional backgrounds. However, it seems that this apparently good result may be an artefact of the procedure of including all the varieties of test types and student types without differentiation. Too much test information has been lost by combining all these groups. It would be revealing to see an analysis of the test results broken down for identifiable (sub-) populations.

Gustafsson (1980; 230) shows how the unidimensionality assumption of the Rasch model makes it, in principle at least, a useful model for the investigation of the dimensionality of a set of items. He reports on the analysis of a test of English grammar for Swedish students where it was found that a set of items measuring knowledge of irregular verbs did not fit the model. But in a separate analysis of these items it was found that they *did* fit the model, as did the rest of the items after some poorly constructed items had been excluded. Had the items measuring knowledge of irregular verbs been excluded, that would have implied that the scope of the test would have to be narrowed unduly, whereas through forming two scales instead of one, both kinds of items were retained. This suggests that there are at least two dimensions to 'English grammar'; what those dimensions might be in substantive

terms would be more difficult to say – it could be, for example, that the 'irregular verbs' dimension is similar to a 'memorisation' dimension. Further investigation would be necessary.

Willmott and Fowles (1974; 31) also give a similar example in the context of the analysis of fit: in an English test they found that one item which asked about the grammatical structure of a phrase in the comprehension passage used had to be rejected on the grounds of misfit. It was said not to be measuring 'English comprehension'; on the other hand it could be said that such a misfit demonstrated the possibility that there might be at least two *distinct* aspects of attainment in English, namely 'comprehension' and 'grammar'.

Haertel (1984) sets out to show how reading comprehension items represent a case intermediate between the two extremes of arithmetic items (where one is applying a learned procedure) and history items (where one is, at least at the lower levels, demonstrating knowledge of a particular fact). Reading comprehension items do require some common set of skills for their solution, so the items can be considered as multiple indicators of the same ability to comprehend text. On the other hand, that ability is far more complex than the simple algorithm required to solve a two-column subtraction problem. Haertel (op. cit.; 65) shows how reading comprehension items in one latent trait analysis fall into two classes: those that conform to the model (and which appear to be largely 'inference' items) and those that do not conform to the model (which appear to be predominantly 'literal comprehension' items).

This is not entirely surprising since no latent trait model could claim to test 'knowledge', but rather the application of knowledge. Haertel (ib.; 70) suggests that it is likely that the difficult inference items could be referenced to the same latent dichotomy because each required the construction of some mental representation of, in this case, the meaning of a paragraph and also required the querying of this structure to solve the problem posed by the item. In contrast, the literal comprehension items were more amenable to various particular, idiosyncratic solutions – for example, choosing a response alternative containing words appearing in the text, or eliminating unreasonable distractors. For testees unable to comprehend the text as intended, the literal comprehension items were more like the example of history items given above – solution of the problem reflects possession of particular, discrete pieces of information, not the exercise of a common skill.

5.1.6. Evaluating fit to the model

Various methods of evaluating fit to the Rasch model are available to us and these will be outlined here. First, however, we should be aware that there are certain difficulties associated with the business of 'fitting the data to the model' or vice versa. The main danger is that we will only accept items (or persons for that matter) who fit into out model and meet the somewhat strong assumptions that are made. Clearly this is a flawed procedure, since it places the test, the use to which it will be put, and the performance of examinees at the mercy of a model which is, at the very least, capable of criticism.

The 'homogeneity' requirement of the Rasch model, for example, is a very restictive condition, which means that in practice it often happens that a considerable number of test items does not conform to the model and has to be deleted. Unfortunately, in many cases no psychological or substantive reasons can be found for some items being conformable and others not. At any rate, one can conclude with Fischer (1978; 301-302) that it makes no sense simply to collect a large number of items and to apply the Rasch model in order to select the homogeneous ones; in such a procedure the test of the model becomes ineffective. The model can, however, be very useful if the item material has been pre-selected cautiously on the basis of a sound psychological hypothesis; then the Rasch model will aid in testing and correcting this hypothesis and provides a well-grounded metric for item and person measurement.

Gustafsson (1980; 220) also makes the point that with very large samples there is a particular problem because we cannot reasonably suppose any model to be perfectly fitted by data, so with a sufficiently large sample any model would have to be discarded. This is essentially a logical objection to using IRT methods (or indeed any testing methods?) in order to be able to claim universal significance. It would appear that the logic of looking for sample-free estimates of ability and difficulty (at least in the 'wonder ingredient' sense which Wright and Stone (1979), for example, claim) is undermined at the beginning.

Wood (1978) has highlighted the problem of fitting the Rasch model by pointing out that the arguments used can often be circular Wright and Panchapakesan (1969; 25) for example say that if a given set of items fit the model this is evidence that they refer to a unidimensional ability, that they form a conformable set, and that fit to the model also implies that item discriminations are uniform and substantial. Wood suggests that not only is this argument circular, but that the only criterion for fit in this view is conformability. In practice, however, test constructors are constantly

faced with the opposing demands of homogeneity and heterogeneity. On the one hand they want to "cover the syllabus" by sampling content according to a specification, while on the other they worry about biserials being above a certain notional figure (thus demonstrating their conformity to the test 'as a whole'). This is, in essence, the substantive argument behind the unresolved discussion between Horn (1968) and Ebel (1968).

Life, of course, is simpler if we believe in homogeneity, and in most practical applications unidimensionality, test reliability, uniform ICCs and fit to the Rasch model imply and are implied by each other. Despite relatively sophisticated measures of fit which are available (e.g. Wright and Stone 1979) it may well be that Wood (1978) is right and that there are so many issues concerning the nature and purpose of measurement that fit to the Rasch model or any other latent trait model is "almost irrelevant". Wood claims that nothing is necessarily to be deduced from items fitting or not fitting.

While not disagreeing with the essential thrust of this argument, we nevertheless need to have some measure of evaluating how our model is performing. Whether we then go on to reject items or persons that do not conform will be dictated by other considerations. But as a first step we need some guidelines to tell us how much confidence we can have in the results that we are obtaining.

The basis of our analysis of fit will be the standardised residual. This takes into account expected and observed responses and allows us to evaluate the quality or significance of measurement error, both for persons and for items.

The standardised residual is estimated by subtracting the expected response probability from the observed response and dividing by the standard deviation of the sample. This set of values is normally distributed with a mean of 0 and a standard deviation of 1. A large negative residual in a person's response to any item means that the person was expected to answer correctly but didn't, whereas a large positive residual indicates that the person was expected to answer incorrectly but didn't. The larger the residual the more unexpected the response. The sum of the squared standardised residuals (Σz^2) ought to follow a χ^2 distribution with 1 degree of freedom for each z^2 minus the d.f. necessary to estimate the person measure (b in the BICAL program); this sum of the squared standardised residuals divided by the d.f. should follow a mean square distribution, and this is evaluated in Rasch modelling as the *t-statistic,* an important measure of fit, which has approximately a unit normal distribution (see Wright and Stone 1979; 70).

There are three basic parts to summarising the information contained in the residual (Smith 1985; 434 -435). First we have the *unweighted total* fit statistic, which squares the residual, standardises it and sums it over all the person interactions. Secondly, we have the *unweighted between fit* statistic, which is a direct test of the 'item-freed' measurement property of the Rasch model. That is, if the data fit the model, the overall ability estimate should accurately predict the person's score on any subset of items. By comparing the person's predicted score with the observed score on any subset of items it is possible to test the fit of the data to the model.

Both of the above statistics are based on the overall ability estimates. To test the fit of a response to the subset of items it is possible to estimate the the person's ability on that subset alone and to create an *unweighted within set* fit statistic for each subset.

The total fit statistic, providing a single overal test of fit, is better at detecting random guessing and sloppiness, as well as other types of random measurement disturbance. The between fit statistic is better at detecting systematic forms of measurement disturbance e.g. plodding, and disturbances resulting from specific item/person interactions. This requires an *a priori* specification of non-overlapping subsets of items for an examination, e.g. grouping items on the basis of item difficulty, position on the examination, item type, specific forms of bias etc. It is possible to perform any number of between tests on each response pattern. The within fit statistic is not very useful in detecting new instances of measurement disturbance. Its greatest utility is in helping to clarify the causes of measurement disturbance detected by the other two fit statistics.

The only difficulty with the t-statistic is that, first, the different t-tests (in an IRT sense) do not necessarily agree, since they offer information about different aspects of fit to the model, as just outlined. This has the advantage of allowing a variety of different perspectives on the analysis of fit, but suffers from the disasvantage that no one test is likely to be conclusive - judgment must, as always, be exercised. In particular, it can be confusing as to the reference values of the various distributions of the t-statistics. Computer programs such as BICAL (Wright, Mead and Bell 1980) include in their output plots of, amongst other things, the values of the various t-statistics; it si a striking feature of such plots that they show the different t-statistics to be more or less uncorrelated.

A further feature of the analysis of fit through the use of squared residuals (not so much a problem, more a procedural decision) is that it is misfitting *persons* that are

omitted from recalibrations, and not misfitting *items*. Misfitting persons are omitted because it may be the responses of the misfitting persons which cause apparent item misfit, so if a test is recalibrated on the basis of rejection of misfitting items then further items will continue to misfit. This, at least, is the argument put forward by Wright and Stone (1979; 81). It could of course be argued that the procedure should be carried out the other way and that the misfitting items should be omitted from recalibration - on the whole, however, the aim is generally to produce as complete a test as possible, and to omit items will reduce the efficiency of the calibration and test development procedure. This decision is already made in the BICAL program, but it should be borne in mind.

Other measures of fit available include the so-called discrimination index, which has a reference value of one: any value greater than one means that the observed ICC is flatter than expected and is therefore failing to discriminate adequately between testees; any value less than one means that the observed ICC is steeper than expected and may thus be discriminating too sharply (in the stepwise fashion of a strict implicational scale). There is a close relationship between the 'discrimination index' and the point biserial (as there is with the traditional discrimination index), and the use of this index depends on a reconceptualisation on the part of the test user to think in terms of ICCs rather than isolated items (see Wright, Mead and Bell 1980; 53).

A final measure of fit which it is useful to think about is the extent to which an item calibrated on the lower-ability group of the calibrating sample agrees with the difficulty value obtained when that item is calibrated on a higher ability group. One would expect the proportion of correct answers in a lower group to be lower than in a higher group, but it is useful to have some measure of how much lower one would expect that proportion to be. In analyses produced by computer, an indication is given of how ICCs calculated for different score groups (up to 6 of them) depart from expected values; this helps us to see if an item is performing uniformly well across the calibrating sample.

To summarise the different measures of fit which can be used and to show how the problem of fit is essentially the problem of how to assess and evaluate the whole test, we now give Hambleton and Swaminathan's list of approaches for conducting goodness of fit investigations (Hambleton and Swaminathan 1985; 157-158):

1. Unidimensionality: plot of eigenvalues of the inter-item correlation matrix; comparison of two plots of eigenvalues (including random data); plot of content-based versus total-test-based item parameter estimates; analysis of residuals

2. Equal discrimination indices; analysis of variability of item-test score correlations (e.g. point biserial correlations, which of course apply to the one-parameter model as much as to the other models); identification of percent of item-test score correlations falling outside some acceptable range

3. Minimal guessing; item-test score plots; performance of low-ability examinees on the most difficult test items; item format and test time limits

4. Nonspeeded test administration; compare variance of number of items omitted to variance of number of items answered incorrectly; investigate percent of examinees completing all, 75%, and 85% of test

5. Invariance of item parameter estimates; comparison of subgroup estimates

6. Invariance of ability parameter estimates; comparison of two or more item samples

7. Checking model predictions of actual test results; investigating residuals and standardised residuals, ICCs and plots of test scores against ability estimates

All of these measures of fit can be useful at different stages in the test evaluation procedure. No one measure is likely to prove satisfactory by itself, since there are always going to be different aspects of the test that we shall want to examine. Nevertheless, the availability of such measures should ensure that our evaluation of the test and its items is based on the fullest possible information.

5.2. Item banking and Computer Assisted Test Construction

Item banking has always had a close connection with computer-assisted test construction (CATC). This is partly because the computational resources needed to analyse large banks of data can really only be handled by a computer, and partly because the uses to which one might wish to put an item bank, as suggested in the introduction, often need a computer to make rapid administrative decisions which are based on complex collections of items.

5.2.1. Defining the areas of interest

Besides the mere scoring of answer sheets and the analysis of data Lippey (1974) recognizes four different areas of CATC:

1. item banking

2. item generation

3. item attribute banking

4. item selection

In fact, these can be more usefully seen as two broad areas, namely *item banking* and *item generation* with item attribute banking and item selection forming a subset of item banking. We shall now consider these areas, for discussion later.

5.2.1.1. Storage

Items are stored in the computer in exactly the same way they are to appear on the test. The computer is then used to select,assemble and present these items to test takers, either in a pre-determined way or in an interactive mode in which the item that is chosen depends on the performance of the examinee on previous questions.

Lippey (1974) suggests that one of the chief disadvantages of item banks in this sense is the remoteness from the user and therefore his lack of personal control. Moreover an adequate classification system that enables each user easily to identify the characteristics of the items he wants is essential.

Millman (1980) suggests that possible advantages are that, since more items are stored in the computer than appear on any one test, it is possible to produce multiple tests having different items arranged in different orders. For Lippey (1974) a shared item bank can satisfy the needs of many users whilelending itself to the efficiencies of centralised management.

5.2.1.2. Item attribute banking

This is really a question of classification; given a bank of items, how does one label them so that a potential test user can select the items he wants?

This has been a major problem with item banking, especially once the bank grows to any size, since for obvious reasons one user's ideas of a test topic might not correspond to another's.

Lippey (1974) suggests two categories of classification:

1. Those item attributes which are dependent upon data collected from item usage, called 'measured attributes';

2. Those which are intrinsic to, or implicit in, the question, or 'assigned attributes'.

The latter will depend upon such notions as subject matter classification,topic headings, behaviourally stated objectives, keywords etc. "Their most valuable function is in selecting the questions desired." (Lippey 1974;9)

5.2.1.3. Item selection

This is closely related to the previous category; the test constructor specifies attributes of the questions desired and the computer retrieves the questions accordingly. It is common practice to specify only a subject matter area and to have relatively few items selected from a large assortment in order to be able to obtain a wide variety of different tests covering the same material.

Problems in this area include, especially with large banks, the questions of copyright and payment (cf. for example Buckley-Sharp and Harris 1970)

5.2.1.4. Adaptive testing

This is not one of Lippey's (1974) categories but is considered by Millman (1980) under the question of item banking.

Adaptive testing is a general term used to describe several procedures in which the particular items administered to the examinee at one point in time depend upon the examinee's performance on items administered at a previous time. A computer is not actually required (cf. Lord 1971), but the interactive feature of between item administration and the individual can be facilitated by use of the comuter.

Millman (1980) distinguishes two types of adaptive testing: in one form the *number* of test items administered to each person varies (Wood's (1973) 'sequential analysis') so that examinees functioning at a level close to the standard are administered more items so that their status with respect to the standard can be evaluated with greater accuracy.

The second form of adaptive testing (Wood's (1973) 'staircase methods') varies the *difficulty* level of the test items administered to each person so that the examinee's true status can be more accurately defined.

According to Millman (1974;36-37) in neither form of adaptive testing is a clearly specified domain used. Depending upon how the population of items was created, the

score derived from such testing procedures might be a good estimate of the student's performance with respect to some construct, like mathematics ability. But without a reference to the specific tasks that are included no criterion-referenced interpretation is possible.

5.2.1.5. Administration

What all these CATC categories have in common is that items are essentially stored in the form in which they will be used and that theoretical interest centres on the classification and arrangement of the items. O'Reilly et al. (1973) consider that the increasing number of attempts to bring computer technology to bear on the testing component in education will have little or no qualitative impact on the functional utility of testing in the learning process, for the reason that the bulk of existing CATC projects take the classification problem as the basis for development and construction and for the use and interpretation of the test data. Consequently the user is asked to do little more than to load in the items which supposedly represent the content of his course. For O'Reilly et al. this lack of effort in determining the intent of testing in relation to other broad decision types results in a "largely irrelevant test development procedure which will deliver tests primarily to maximize score variance among students [with] overreliance on machine processing." (p.33)

This seems unduly pessimistic; a consideration of classification types will surely help to focus on tighter definitions of course content, and is hardly a trivial question in the case of adaptive testing. To say what a test is a test of has always been of central importance.

5.2.2. Item Generation

A far more radical use of CATC than the simple administration of item banks is the generation of test items. This is accomplished by developing a set of rules (an algorithm) which, when carried out by the computer, will result in a complete question. The specific item produced, although it belongs to a known domain, is unknown until it has been generated.

The algorithm may be considered as a model for the question, as a framework which defines its broad properties, or as a question skeleton which requires further specification to become a complete question.

Lippey(1974) points out that the value of item generation stems from the fact that one algorithm can be used to produce a large number of questions and that "... any

discipline which follows well-defined rules lends itself to item generation."(p.8) Many of the technologies and techniques to be discussed in chapter 4 are of this 'algorithmic' type and therefore suitable for CATC (cf., for example, the work of Fremer & Anastasio 1969 and Berk 1978)

Each item program represents a small domain of items, and item programs rather than the items themselves are stored in the computer. To make a test, one or more item programs are executed one or more times each. Because each item program is capable of producing a large number of variations in the content of a specific item, it is unlikely, as Millman points out (1974;38) that any two given tests will contain exactly the same items. The advantages for any situation where test security is a problem are obvious.

5.2.2.1. Quantitative Items

The problem is that this type of algorithmic test construction is best suited to quantitative items. Prosser (1974) emphasises the role of the 'prototype' or model of an item's structure and claims that the successful generation of test items through automatic processes requires that the rules for creating specific entries for the variable parts of an item prototype must be completely and concisely specified; for him, algorithms that require later human editorial judgment are apt not to be of value.

A glance at what has been done in this field would certainly seem to bear this out. Hively et al.'s (1968) item forms technology, while not specifically designed for computer implementation, was concerned almost exclusively with arithmetic items. Hsu and Carlson (1973) attempt a more thorough-going attack on the problem, though still within the field of mathematics. As part of an Individually Prescribed Instruction package they analysed the objectives contained within a unit and divided each objective into several item forms, established by "logical analysis and by examination of existing test material to determine the types of errors students tend to make within each objective." (p.26) For them, the construction of item forms is the most important factor in determining the validity of the test items.

For our present purposes we should note

- the relation of the test construction process with units of instruction. This is similar to the problem of specific and generic CAL; and
- the emphasis on content validity at the item-form construction stage.

5.2.2.2. Non-quantitative items

As far as the use of this technology for non-quantitative items is concerned, various attempts have been made, though they suffer from apparently unavoidable limitations: the tension between freedom and restriction is difficult to resolve.

The problems encounted by Fremer and Anastasio (1969) have already been discussed in chapter 4.

Richards (1967) attempted to determine whether, in principle, some tests for screening college applicants could be written by computers; 'in principle' because the test he chose was not actually written by computer but by a proposed algorithm which a computer might use. Concentrating on verbal comprehension as the test factor most useful for predicting academic success in college, Richards developed a procedure for writing synonyms based on word classifications according to Roget's *Thesaurus*. Richards (1967;211) points out that the most difficult problem in writing tests on a computer is developing a sensible procedure for choosing distractors in multiple- choice questions. This is a concern echoed by many others in the field (e.g. Prosser 1974;64) and highlights the need for pedagogical insight rather than exhaustive listings of potential items as a criterion in item construction.

Richards (1967) compared his 72-item computer-written (or 'writeable') synonyms test with the *Wide Range Vocabulary Test* (WRVT) from the ETS *Factor Kit*. The items in the computer-written test turned out to be easier, on the whole, than those in the WRVT and the reliability of the computer test was lower, a disappointing result in view of the greater length of the computer test. On the other hand, the fact that the computer-written items were completely randomly generated suggests that in principle the method is workable, and indeed Richards (1967;214) concludes that existing computer technology would enable many if not all aspects of college admission testing to be automated.

Clearly, synonyms tests of this sort have their place in L2 testing, but whether one would wish to concentrate one's efforts on automation at the expense of a more serious investigation into the actual use of vocabulary is open to doubt. The main question seems to be how to define adequately the domain being investigated.

5.2.2.3. Sentence generation

More usefully as far as L2 reading is concerned, attempts to generate items based on the sentence would seem to offer hope. Most such attempts, however, have been concerned with specific content areas, and test knowledge rather than language. In

certain circumstances, however, it is highly likely that one would wish to employ such techniques - ESP situations for example.

Vickers (1973) generated questions concerning the FORTRAN programming language in a random but guided fashion. The student was asked to recognize various valid and invalid FORTRAN language elemental constructions. Each test contained 50 items, chosen randomly from a set of 5 possible types of construction. The major disadvantage as Vickers saw it was that the student need know only how to recognize the various elements of the programming language and is not required to understand the 'semantic meaning' of the various constructions. This is only a problem if one does not wish to test recognition only, which may be the case with computer programming, but not necessarily so with other subjects.

Denney (1973) tries to combine the virtues of item banking (flexibility which comes from the storage of enormous volumes of data in fixed format for rapid retrieval) with those of item generation (where a few complex algorithms can generate large numbers of 'relatively different' questions). He sees the costs of the former as being in storage, those of the latter in program logic and computation. Denney's (1973) Question Pool Management System (QPMS) depends on the identification of three parts of a multiple- choice question: the question stem, the potentially correct answers and the incorrect answers.

QPMS permits an association of up to 7 correct answers and 7 distractors with each question stem, so if one assumes that only one of the choices will be correct there are 245 ways that a choice can be made of one of the seven answers followed by four of the seven distractors. Since QPMS also randomises the arrangement of the 5 choices there is a total of 29,400 various physical formats that can be generated from a single question. Now, this may not be entirely suitable for L2 tests; Denney's (1973) system seems to be little more than a sophisticated substitution table.

Olympia (1975) extends Denney's (1973) system somewhat, claiming that the construction and scoring of examinations are primarily clerical functions; nevertheless, he has developed a versatile algorithm for chemistry, maths and physics, which can generate repeatable examinations from a compact data base. For all the multiple choice items there is a keyphrase pool, a statement phrase pool and a distractor pool (e.g. "Oslo (...) is the capital of Finland (...)") Significantly, however, thesystem is closely associated with specific course material, since the data bank is arbitrarily divided into sections corresponding to each unit or block of a course associated with specific course objectives. Were it possible to agree upon the suitable 'pools', such a

system could readily be adapted to L2 testing.

The type of technology to be discussed in chapter 4 for linguistic transformations seems to be the most promising and interesting aspect of future developments in item generation (see, for example, Bormuth 1970, Anderson 1972, Finn 1975, and Royd & Haladyna 1982).

5.2.3. Adaptive Testing

The problems associated with diagnostic testing, item banking and adaptive testing are, to a large extent, shared. The essential issue is how to ensure that an item has validity such that it is applicable to a wider group than some sort of normative population. There is, therefore, a common concern with 'latent traits' and the measurement of 'ability' in relation to item difficulty. We shall look at the question of adaptive testing to see what relevance it is to item banking.

What is of interest here is that adaptive testing represents what Urry (1977) has called a 'remarkably effective application of latent trait theory' and that the reality of adaptive testing is the fruition of the "inevitable computer conquest of testing". Urry sees adaptive testing as best implemented as a computer-interactive method, so that the test is terminated when the estimate (of ability) reaches a spefified level of reliability. On the other hand, even accepting a latent trait model poses difficulties in that tailored testing becomes less effective when a model with insufficient parameters is used (Urry 1977;184). Indeed, Urry found that the Rasch model was found to be particularly inappropriate for use with multiple choice questions because the model (a 2-parameter one) did not portray multiple choice response data with fidelity. This question of models for multiple choice questions has already been discussed and will not be repeated here.

5.2.3.1. Terminology

All adaptive testing involves selecting for administration to a given individual the set of test items, from all the items available, that is likely to measure that individual best (Weiss 1980;137). More particularly, it represents a family of testing systems in which the choice of items is governed by the subject's response to item (n-1), or, in some cases, his responses to items (n-k) to (n-1) where k is a small number (Wood 1973;529).

Other names which all describe the same phenomenon include:

- programmed testing
- branched testing
- sequential item testing
- tailored testing
- dynamic testing
- response contingent testing

These terms will be used interchangeably here.

5.2.3.2. Basic ideas

The basic idea of adaptive testing is that given some initial information concerning a person's ability, reliable or otherwise, and with the aid of a computer, that person be presented with an item designed to effect the greatest reduction in the uncertainty about his true ability - one for which he has a 50% chance of success. Having observed the outcome, the estimate of ability is then revised - upwards if he got the answer right, downwards if he got it wrong. Another item is then chosen with the same optimal properties as before and the cycle is continued until the uncertainty is resolved satisfactorily.

Ideally, everyone is agreed, adaptive testing is best implemented in an interactive computerised set-up. But as Wood (1973;530) points out, only when the pay-off is likely to be high will the expense be justified.

5.2.3.3. 'Ability' and 'difficulty'

It will be seen that adaptive testing aims to assess 'ability'. As such it has little to say which will aid specific diagnosis.

The problem, as with item banks in particular and latent traits in general, is to define 'ability' (and item 'difficulty'). Wright (1977) puts the case for defining ability and difficulty in terms of performance on test items: "Do we want to think that the more able (person) has a better chance for success no matter what the difficulty of the attempted item? Is that what we intend 'more able' to mean? Similarly, if we want to think that the probability of success on the harder of two items should always be less than the probability of success on the easier, no matter who attempts the items, if that is what we intend by 'harder', then we must see to it that variation in item discrimination sufficient to produce item characteristic curves that cross does *not*

occur." (p.103)

Weiss (1980) tries to assess the importance of other referents. Firstly, there is the relationship between ability testing and achievement testing. In general both are concerned with measuring an individual's performance on a certain type of task. Generally, when a test is used to measure how well a person can do one of these tasks during or at the end of a specified period of instruction, the test is an *achievement* test. The same test items can be used to measure a person's *ability* when the capability to perform the task has developed over a long period of interaction with a variety of environments (Weiss 1980;131).

Thus, ability is the capability of performing a task that results from an unspecified learning history over a long period of time; achievement may be the capability of performing the same task as the result of specific instruction or training.

Such considerations, along with the awareness of *population + content area + point in time* (e.g. knowledge ofgeneral biology might be tested to an 85% criterion level at the end of 3 weeks' instruction) leads Weiss (1980) to a 'multivariate' conceptualisation, which considers achievement as occurring differentially within content areas for a given student.

Thus, for Weiss, the criterion-referenced (or content-referenced, or diagnostic) problem should be to determine whether or not an individual has mastered not only one content area at one point in time, but a number of content areas, possibly at different levels of mastery.

Thus, says Weiss, "it is important in the criterion-referenced measurement of achievement to consider mastery as occurring separately for each identifiable subdomain within the course. At the same time, an individual's achievement levels should be monitored with respect to both time and populations (i.e. different mastery criteria). The problem of achievement measurement, therefore, becomes one of simulataneously and continuously measuring an individual's achievement levels on the multivariate, multi-time, multi-population cube."(p.137)

5.2.3.4. Adaptive testing methods

A number of adaptive testing methods are available.aSEQUENTIAL ANALYSIS. In the sequential analysis system, a subject is presented with a sequence of items, ostensibly of equal difficulty. Indeed, the items are supposed to be wholly equivalent in the sense of being exchangeable. If the proportion of correct responses exceeds a

predetermined value, testing ceases and the subject is allocated to an appropriate treatment; and similarly if his correct response rate drops below the target figure. In the event that neither decision can be made, testing is continued.

In this method the nature of the n^{th} item is not contingent on the $(n-1)^{th}$ response, for all items are supposed to be identical. The only thing that is contingent is whether there will be an n^{th} item. This seems to offer little of any real value. Moreover, there are uncertain practical problems here: to insist that items should be identical is unrealistic. Wood (1970) and Ferguson (1969) argue that sequential analysis methods are well suited to criterion-referenced testing, where it is understood that resources are not inexhaustible and that something less than 100% mastery must be accepted at some point. Using item form techniques Wood (1970) hoped to generate equivalent items from the relevant universe, but "the problem of defining non-trivial universes remains vexatious."(Wood 1973;531)

'STAIRCASE' METHODS. Patterson (1962) took a pool of items and arranged it in such a way that a person starting with an item of average difficulty would, on encountering an item and getting it right, proceed to a harder item; otherwise he would proceed to an easier item. The subject thus traces his way through a network of items.

Linn et al. (1969) compared seven programmed tests of this type; against the criterion of reproducing the 190-item total test score in the cross- validation sample the programmed trests were found to be only slightly superior to the shortened conventional tests. However, the programmed tests had correlations with the outside criterion tests that were substantially higher than the corresponding shortened conventional tests. It was estimated that a test which was parallel to the 190-item total test would have to be 3.36 times as long as the best programmed test to have an equal median correlation with the outside criterion tests.

STRADAPTIVE TESTS. A stratified adaptive, or 'stradaptive', test (Weiss 1980;137) assumes that all the items in the pool measure a single dimension. All the very easy items are combined into a subgroup called a 'stratum' and so on. The stradaptive test operates from a pool of test items that are stratified by difficulty levels (with an up-one-stratum branching rule for correct answers). Weiss found that measurements provided by the full-length adaptive test were of considerably higher precision at all achievement levels than those of the conventional test, even though the adaptive test was, on the average, 3 times shorter than the conventional tests.

FLEXILEVEL TESTS. Lord (1971) maintains that a conventional test becomes a flexilevel test when modified so that the examinee follows these rules:

1. Answer first a specified item of median difficulty;

2. After answering an item correctly, attempt next the easiest unanswered item of more than median difficulty; after answering an item incorrectly, attempt next the hardest unanswered item of less than median difficulty.

This uses a special answer sheet (the test is not computerised) so that the examinee knows whether each answer is correct or incorrect. If the conventional test contains N items, the examinee taking the flexilevel test will attempt only n=(N+1)/2 of these.

Lord (1971) found that with examinees of average ability levels the standard test was quite adequate, but for good discrimination at the extremes of the ability scale, then a flexilevel test was better.

BRANCHED PROGRAM ACHIEVEMENT TESTS IN L2. Boyle et al. (1976) and Walton et al. (1979) describe the development of a branched program achievement test (BPAT) for French, which is claimed to have strong diagnostic possibilities. Although the project is said to be 'computer-mediated testing', in fact only the scoring is done by computer, and the actual test shares many of the characteristics of Lord's (1971) flexilevel pencil-and-paper tests. The test is written to include multiple questions of each of the larger concepts to be evaluated (with the fields of: 1. verb morphology 2. grammar 3. vocabulary) The BPAT was course-specific and differs little from a jumbled conventional test (cf. e.g. "La lionne n'est pas une _____" bete 94 chatte 03 chienne 41. The numbers refer to continuing frames)

5.2.3.5. Problems associated with adaptive testing

Lord (1971) wonders if flexilevel testing will be too confusing or too time-consuming for many examinees. This would certainly apear to be the case with Walton et al.'s (1979) BPAT. Weiss (1980) suggests that there are two reasons why many of the adaptive testing strategies developed for single-content-area ability tests may not be appropriate for achievement tests that cover several content areas (or specific subdomains of achievement): the first reason is that although the unidimensional branching models can be applied to separate content areas, they are not designed to take into account the information available between content areas.

The second, more practical, reason is that it might not be possible to generate relatively large numbers of items within one content area in an achievement test. Lord (1970;179-180) shares this view. He points out that if, for example, 500 itemsare available for tailored testing, better measurement will often be obtained by selecting

the 60 most discriminating items and administering them as a conventional test rather than by using all 500 in a tailored testing procedure. This, he says, "may actually prove to be a fatal objection to any general use of tailored testing." He further suggests that even supposing the items in a pool can be assembled into a branching structure, the longest tailored test that can be drawn from a 500 item pool is 31 items (using the formula quoted on page 143 above). The superiority of a 60-item test with highly discriminating items is thus not hard to believe.

Wood's (1971) attempt to get round this using a Fully Adaptive Sequential Testing(FAST) procedure, where the step size did not reduce regularly according to a predetermined formula but instead was adaptive to the item responses as they were made produced ambiguous results. FAST achieved considerable reductions in measurement error, but only in some parts of the ability range. Wood concludes that no one testing method enjoyed across-the-board superiority; the choice of the most efficient testing method for an individual depends to a large extent on local knowledge,both about the person 's ostensive ability and the quirks of the instrument, particularly that region of the item pool where he will be expected to encounter most items.

5.2.3.6. Evaluating adaptive testing

One has to be careful in evaluating adaptive testing. There are certain areas of knowledge where the precision and accuracy of test results are critical; this is generally not the case with L2 learning. We might evaluate adaptive testing according to three functions it can serve (Wood 1973;538-539):

1. As a means of individualising group tests;
2. As satisfying measurement needs which are beyond the scope of conventional test instruments;
3. As a substitute for orthodox,individually administered tests.

Much depends on the measurement requirements; if the purpose of testing is to select, say, the top 25% of a population for further education, then the need is demonstrable. "The point, as always, is to appraise the measurement situation and decide what is the best instrument to use in the circumstances, given the resources available." (Wood 1973;539)

5.2.4. Test equating

5.2.4.1. Introduction

The field of test equating is one that has an intimate connection with item banking, indeed many of the concerns of the two fields are such that in many circumstances the two may be considered as slightly different perspectives on the same problem (see e.g. Hambleton and Swaminathan 1985). The reason for the importance of test equating is that it is directly concerned with major problems in test development. In the case of large-scale testing in particular it is necessary to know how to equate test scores on different versions of a test.

For standardised testing programmes, the use of reliable and valid tests requires that such tests be revised at regular and frequent intervals. For purposes of continuity it is necessary to have a scoring system whereby scores on newer versions can be compred directly with scores on earlier versions. This equating procedure adjusts for differences in test and item characteristics.

Test equating is, conceptually, a fairly simple idea, but its application to actual test data is full of practical problems and complications. These include such issues as omitted items, guessing, unequal test reliabilites, and different scoring schemes (Potthof 1982). There are two main forms that test equating can take: first, the forms to be equated are designed to be as psychometrically identical as possible. They are intended for the same examinee population. The tests are written to measure the same 'ability' at a comparable level of difficulty. This is the most common type of test equating procedure and is often referred to as horizontal equating.

The second form of equating is referred to as vertical equating (though the American Psychological Association 1985 suggest using the terms *scaled* or *comparable* scores rather than *equated* scores for most situations). For vertical equating, the tests to be equated are intentionally different in difficulty. The purpose here is to linkk tests that measure an ability over a very wide range. It is in this linking of tests over a wide ability range that we see the connection with item banking. Moreover, the fact that in vertical equating one may be interested in performance on tests over time (perhaps asking children to retake certain tests at various stages of their school careers) means that the dimensions of the test are particularly important: is, for example, the material that is to be tested in the fourth year of school in some sense on the same scale (in terms of dimensionality) as the material that was tested in the third year?

The fundamental issue (and it is in essence the issue at the heart of item banking, hence the criticisms of Goldstein , Tall and others – see later in this chapter) is how well we can measure human growth (Skaggs and Lissitz 1986).

5.2.4.2. Equating methodology

One may define equated scores as follows: Two scores, one of Form X and the other on Form Y (where X and Y measure the same function with the same degree of reliability) may be considered equivalent if their corresponding percentile ranks in any given group are equal (Angoff 1984; 563). Under this definition, two notions are included: first, Form X and Form Y measure the same trait. Second, the equating relationship must be unique. This in turn implies that the equating is independent of the samples used to derive the equating and that the two tests are equally reliable and parallel.

A variety of equating methods have been developed over the years. Most of these can be grouped into two general classes of models: linear and equipercentile equating. The definition given above, which is based on quantiles, is a direct expression of equipercentile equating. In equipercentile equating one attempts to form an agreement between all moments of the raw score distributions on the two tests to be equated. If the shapes of the distributions differ the equating function will be curvilinear.

Linear equating, on the other hand, assumes that all moments beyond the second are equal in the two score distributions, that is they have the same shape.

Given this assumption, raw scores on two tests can be considered equivalent if they correspond to equal standard scores in any given group.

In practice, linear equating is often preferred. It is both simple and analytical, requiring only the calculation of a linear function to transform scores from the scale of one test to that of the other. With equipercentile equating one usually wishes to smooth the cumulative distributions, and this often introduces some subjectivity into the process. For vertical equating, linear equating is clearly inappropriate since the raw score distributions will differ in shape (i.e. moments beyond the second will differ) for samples of equal ability. A full discussion of the practice of conventional equating can be found in Angoff (1982, 1984), Braun and Hollan (1982), Flanagan (1982) and Potthof (1982).

5.2.4.3. Experimental design

The design of studies for test equating shares much in common with the design of studies for item bank construction. There are three basic equating designs. In the first design, two random samples of examinees are selected and each is administered one of the two tests to be equated. Because the samples are randomly selected they are expected to be equivalent in ability. In the second design, both tests are adminstered to a single sample, with the order of the tests counterbalanced to control for practice and fatigue effects. In the third design, there are two samples: one test is administered to each group and an anchor test is administered to both groups. This design, called an anchor test design, is the one most commonly found in real situations (Skaggs and Lissitz 1986; 504). The anchor test may be internal, that is it may form part of both tests, or it may be a separate external test. The samples may be non-random, intact groups, and the anchor test is intended to adjust for any between-group differences. The anchor test design is also the most complicated, but constraints on large testing programmes often require its use. A fuller discussion concerning anchor test design follows later in this chapter.

5.2.4.4. Equating studies using the Rasch model

The first studies using the Rasch model investigated the 'invariance' properties of the model in one of two ways. First, two samples were administered the same set of items and the two sets of item difficulty estimates were compared. This was an example of person-free item calibration. Second, two sets of items were given to the same sample, and the two sets of ability estimates were compared. This latter situation was referred to as item-free person measurement.

Wright (1968) illustrated both aspects of invariance with item response data from the Law School Admissions Test. Test characteristic curves based on high and low ability estimates were very close. The concept of a 'standardised difference' score for each individual was developed at this time too: such differences should have a mean of zero and a standard deviation of one if only random error were present, which was the case in this study (see chapter 7 for a discussion of the standardised residual in relation to language test data).

Anderson *et al.* (1968) found for two samples of nearly equal ability a correlation of .96 between the two sets of item difficulty estimates. However, Tinsley and Dawis (1975) obtained unclear results when investigating four types of analogies tests and samples from four very different populations (correlations between sets of difficulty estimates ranged from -.08 to .98). One possibly contaminating effect was that this

study dealt only with very small samples and short tests. On the other hand, correlations between ability estimates for identical raw scores were in all cases .99, thus suggesting an independence between test and item calibration.

Removing misfitting items had a similarly unpredicable effect: in Anderson *et al.* this resulted in an increase in the correlation between item estimates. In Tinsley and Dawis changes were inconsistent.

Whitely and Dawis (1974) also compared different sets of items. Two ability estimates were obtained for each testee in different ways: odd and even items, easy and hard items, and random subsets of items (resulting in two 30-item subtests in each case). For the odd-even and random set comparisons the means and standard deviations of the standardised difference statistic were very close to zero and one respectively. However the varance of the standardised differences in the ability estimates from easy and difficult items was significantly greater than one. Poor fit to the Rasch model may have influenced the outcome (57% of the hard items and 23% of the easy items did not fit the model).

Loret *et al.* (1974) differed from the above studies in that it is solely concerned with test equating and not the examination of the model. This was a large-scale equating of several forms and levels of seven published reading test batteries which, though not using IRT itself, provided the data for several later studies.

Rentz and Bashaw (1977) was the first of these studies. The outcome was the National Reference Scale for reading, which in effect treated the 2,644 items 28 test/level combinations as a calibrated item pool, any subset of which could produce a score on the National Reference Scale. The adequacy of the procedure was assessed by looking at the variability of ability estimates across repeated administrations of the same test. THe standard deviation at each raw score ability level was computed and averaged for all raw score groups to provide a single index for each test/level combination. These ranged from .008 to .041 logits. Relative to the ability scale itself and to the standard error of an individual's ability estimate, these values were quite small. Therefore the authors concluded that there was sufficient invariance to justify the Rasch equating.

Some of the difficulties identified above, notably the question of the invariance of the ability estimates when subtests are deliberately different in difficulty and the questiron of the potential problems of small sample sizes and samples that are widely different in ability, have been investigated by Slinde and Linn (1977,1978). Slinde and Linn (1977) found that vertical equating using the equipercentile approach resulted in

large discrepancies in grade-equivalent and scaled scores for the same examinees based on different levels of a published test.

Slinde and Linn (1978) investigated vertical equating with the Rasch model in a study that replicated and extended part of Whitely and Dawis (1974) and Wright (1968). Whitely and Dawis had noted that the variance of the standardised difference scores on easy and difficulty subtests was slightly larger than would have been expected with purely random measurement error. This was replicated by Slinde and Linn. In addition, they investigated the stability of equating when item difficulty estimates were derived from one sample and then applied to a different sample. Based on the standardised difference statistic, high ability examinees received comparable ability estimates from the two subsets (easy and difficult) when high ability examinees were used in the equating. Similarly, equal ability estimates were observed for low ability examinees when the equating was based on low ability examinees. This confirmed findings from the earlier studies. However, when a group other than the one for whom the results were applied was involved in the parameter estimation procedure then substantially different ability estimates were obtained from the two subsets (by as much as 1.2 logits). In other words, an examinee's equated score on different levels of a test varied depending on the ability level of the sample on which the equating was based. This was a clear violation of the Rasch model invariance. In defence of the model it can be said that the comparisons involved were severe, the raw score means for the easy and difficult subtests differing by as much as two standard deviations and the high and low ability groups differing by about 1.8 logits. Nevertheless, limits to invariance clearly exist.

Slinde and Linn (1979) generally supported the earlier study with data from the Anchor Test Study, using three ability and difficulty parameter estimates (the middle group was included this time). In particular, widely different ability estimates were obtained whenever the low group was used either for calibration or comparison. The conclusion was that guessing could play a role in the poor results, a conclusion supported by Gustafsson (1979) who criticised the 1978 study by Slinde and Linn.

Using less extremely different groups, Loyd and Hoover (1980) still found evidence that the equating between any two levels (of which they had three) was influenced by the group upon which the parameters were based. No definite trend emerged, except perhaps that an examinee would receive a higher ability estimate if he or she took the same test level as the calibration group. On the other hand, a factor analysis of the total item set (to investigate for dimensionality) suggested that more than one factor was present in the item set. One implication of this study is that certain types of

test, such as curriculum-based tests, cannot be equated because their content changes from level to level. Similarly an item bank could not be constructed for the same reason.

Equating bias was revealed to be present in vertical equating by Divgi (1981) who found that low and high ability examinees would receive a higher equivalent score if they took the difficult subtest rather than the easy one (from the Intermediate Level of the Reading Test of the 1978 Metropolitan Achievement Tests). The Rasch scale favoured medium ability examinees who took the easy subtest. Apart from the inadequacy of a 'once-for-all' approach to Rasch vertical equating, this study also demonstrates the need for caution in the use of the standardised difference statistic as the sole means of evaluating equating results (since the mean and standard deviation of th standardised difference statistic were -.037 and 1.06 logits respectively - a seemingly adequate picture).

Guskey (1981), while not concerned with the above issues of cross-validation, demonstrated that Rasch ability estimates for the ITBS reading test may be more indicative of the actual abilities of the examinees than the publisher's grade-equivalent scores. Guessing was, however, minimised, since only high ability examinees were used.

Forsyth *et al.* (1981) investigated item and person invariance in relation to data that violated the Rasch model to some extent. The conclusion was that the Rasch model yielded reasonably invariant results; however, they did note that the degree of invariance seemed to be related to the difference between average discrimination values for two sets of items.

Holmes (1982) used test items which were deliberately chosen to fit the Rasch model on the assumptions of unidimensionality and equal discrimination and found, perhaps surprisingly, that the results obtained by Slinde and Linn (1978,1979) were replicated - students at different grade levels received higher ability estimates from the test designed to match their grade placement, the largest differences occurring for students in the lower ability range. Guessing - or rather the failure to account for it - seems to be a primary factor in these results.

Using the Rasch model for vertical equating is, then, problematical. Failure to account for chance scoring is probably a major reason for its ineffectiveness. It is not, however, really understood how violations of assumptions affect Rasch equating. Furthermore, the use of the standardised difference statistic may mask error at different points along the range of the ability scale (Divgi 1981). One recommendation

has been that the Rasch model be used in horizontal equating applications but not in vertical equating (Skaggs and Lissitz 1986).

5.2.4.5. Equating studies with other IRT models

The difference between equating studies with the Rasch model and other equating studies is largely a difference of research orientation. The studies using the Rasch model are largely concerned with the stability of parameter estimates, whereas studies using three-parameter and other logistic models are usually concerned with comparing different methods of parameter estimation (usually the three-parameter versus Rasch models).

Marco *et al.* (1983) was one of the first comparative studies of this type. Using data from the Scholastic Aptitude Test they found that, for horizontal equating, if the anchor test was external (see above) and equal in difficulty to the two total tests, both the linear and IRT methods perfomed well. With an internal anchor test, the equipercentile approach also worked well. The Rasch model results were slightly better than those of the other methods when an external anchor test was used. With a parallel anchor test (i.e. equal in difficulty to the two total tests) the type of sample mattered very little. When the anchor test was easier or more difficult than the total tests, equating with random samples showed very little error. On the other hand, IRT models were much superior to the traditional methods with dissimilar samples (samples of unequal ability). Neither IRT model was clearly superior to the other.

When tests of unequal difficulty were equated, the best linear method displayed the largest total error, followed by the Rasch model. The three- parameter model equatings contained the least amount of error when IRT-based criterion scores were used. The equipercentile method was the best method when an equipercentile criteriton score was used. This indicated some bias in the criterion, which means that the results are difficult to interpret.

This study suggested that IRT methods were superior in horizontal equating when samples are not randomly chosen. For vertical equating, the Rasch model produced a large total error (thus replicating the earlier findings reported above). The three-parameter model was far superior to the one-parameter (Rasch) model for vertical equating. This is not entirely surprising if the assumption is made that it is the guessing factor which accounts for the poor performance of one-parameter models, since the SAT-Verbal items are known to be fairly difficult (and thus will lead to more guessing among more test-takers).

Kolen (1981) examined a number of IRT models as well as a linear and an equipercentile method.In this study the tests to be equated had no items in common and each test was administered to an independent random sample. The assumption was that ability distributions would be the same since the samples, including an independent cross-validation group, were randomly assigned to their tests. A complex interaction was seen to exist between item content, difficulty level, and the test model. For vertical equating, the linear and Rasch models performed relatively poorly (a familiar pattern by now), but for horizontal equating the results were inconsistent.

An exception to the general pattern of superiority for three-parameter models for vertical equating can be found in Patience (1981). Using three levels of the total test (the 'Expression' subtest of the ITED) with six overlapping items between adjacent levels (i.e. an internal anchor test design) and 1,000 examinees at each level, Patience found that the three-parameter model was outperformed not only by the Rasch model but also by the two-parameter model and by the equipercentile model (based on correlations between equated and observed scores). Reasons for this apparently unusual result may be that the anchor test was too short, the sample sizes were too small, and the test lacked unidimensionality.

Different aspects of the equating problem have been investigated in connection with equating methods. Sample size as an independent variable was studied by Cowell (1981) using results from TOEFL. The tests were probably very similar in difficulty and the samples were probably nearly equal in ability . Stable three-parameter model equating results were obtained using the small samples. Discrepancies resulting from using small as opposed to large samples were less than discrepancies resulting from using the one- as opposed to the three-parameter model.

On the other hand, Kolen and Whitney (1982) using the General Educational Development Tests found that with small samples (170 - 198) a number of extreme item parameter estimates were produced with the three-parameter model. It seems likely in these studies (Kolen and Whitney (1982) and Patience (1981) that the data fit the three-parameter model to varying degrees.

In their review of the research to date, Skaggs and Lissitz (1986) suggest that the results so far have not demonstrated the consistent superiority of the three-parameter model over other IRT models and conventional methods. The likely confounding factors are parameter estimation problems, differences in data collection designs, different linking processes and multidimensionality.

So far as the anchoring question is concerned, Loyd (1983) has examined the problem of internal and external anchor tests and her results suggest that the external anchor provided more satisfactory results than a series of internal anchors. Moreover, the Rasch and three-parameter models produced quite different equating results, especially for lower ability values. This confirms previous results on vertical equating, but also shows that the equating design can likewise markedly influence equating results.

A potentially major source of problems for IRT equating involves violation of the unidimensionality assumption. Whereas a single dimension is implicit in any test equating, IRT methods might be less robust to this assumption. Clearly this is a substantive issue which needs careful consideration at all stages of test design and analysis.

5.3. Themes of item banking arising from CALL

We should remember that use of any aspect of CATC does not necessarily entail commitment to any CAL philosophy. As Libaw (1974) puts it: "...using computers to assist in test construction has no intrinsic relationship to the pedagogical point of view for which the tests will be used."(p.173). Indeed, CATC is for Libaw a liberating activity which, while permitting the retention of 'slop' in the teaching process itself and allowing the student to select his preferred modes of learning, at the same time opens up the possibilities of 'teaching to mastery'. By this, Libaw appears to mean that CATC provides 'objectives' which teachers/students arrive at as best they can and that mastery learning in general and computer-based learning in particular encourages the breaking down of larger units of instruction into smaller units for which learning objectives can be precisely stated.

It is clear in this case that CATC may reflect or encourage a particular pedagogical conception but is itself silent as regards pedagogical practice. In this respect it is no different from any other form of testing.

5.3.1. Generic versus Specific CALL

An important question to raise in the context of CALL is that of how generic or specific to make the learning/testing package. There must be, argue Kenning and Kenning (1983;11), a question mark over the usefulness of any package which is not co-ordinated with a textbook. A series of grammatical exercises, for example,stands little chance of being effective if it contains many lexical items unknown to the learner. To get round this problem, writers sometimes incorporate an optional section

on vocabulary, but this is very much a makeshift solution.

Culley (1984) defines the two approaches as follows: Specific CAL is that in which a designer focuses on one subject and one textbook. It is characterised by 'hard coding' of questions and answers in a format which is closely and exclusively tailored to one application.

Generic CAL, on the other hand, attempts to reach a wide range of users who may be at different levels in language study and who may be using different textbooks. The critical weakness of generic programs, according to Culley, is that in trying to fit all approaches they fail to provide much real instruction; different textbooks employ differing vocabulary and varying approaches to grammar and syntax. Even where the same vocabulary items are found they are often introduced in a different order. This problem of order applies to grammar as well, so that "... as long as *any* material in the target lanaguage appears in the CAI program, compatibility remains a problem." (Culley 1984;184).

The attempt to write a generic program breeds frustration,says Culley; anything specific or really helpful to the student (i.e. something that provides diagnostic information) limits the usefulness of the program with some textbooks and instructors. He illustrates the problem with Earl's (1981) program: "Alicia: A Spanish Bilingual Reader". Here studentsread a few lines of the story in Spanish and then respond to questions about it. Text and questions may be presented in either Spanish or English. But a student using this program soon runs up against the problems of vocabulary, grammar, and syntax. In effect, a student who knows enough Spanish to make progress through the program does not need the help it offers. Diagnostic information can be provide, but must be limited to a list of language items the student does or does not know (for whatever reason).

Attempts to resolve the problem of the fact that the specific program fits only one textbook (or indeed, often only one small portion of one textbook) have been largely characterised by a tendency to assume the problem does not exist: Comite and Russell (1982) have a program called 'Micro-Deutsch' which presents the material they consider appropriate and assumes that teachers will find ways to incorporate it into the curriculum. Such a program will be successful in so far as it corresponds to what teachers actually do in the classroom.

5.4. Item Banking in Practice

5.4.1. Existing systems

Item banks to date, where they have not been variations on the 'item pool' theme (as is largely the case with the American 'banks'), have been concerned with two basic issues: the first issue is the calibration of items themselves so that a more or less stable bank of items can be produced which have known psychometric properties. This has been the line of approach adopted by most mainstream work in item banking, beginning with Wood and Skurnik (1969) and culminating in the work done by the NFER (exemplified in Willmott and Fowles 1974). The second main issue (though best thought of as an extension of the first) has been with the interpretation of scores/measures of ability which are obtained through the use of item banks. This wide field includes the very large area of vertical equating of tests discussed earlier,but also includes adaptive testing and the establishment of convenient scales of use (e.g. Haksar 1983). A further technical issue has been the use of partial credit scoring (cf. Masters 1984) and other scoring procedures (cf. Pollitt and Hutchinson 1987) which do not require dichotomous scoring.

Since the substance of the main issues has already been discussed, these will not be repeated here. Item banks are responses to specific situational demands and the problems already discussed are applicable in all cases. At the level of the individual bank it is the content matter which becomes important, since it is the psychology of the subject being studies which affects the interpretation of the data and the method of procedure (see e.g. Gustafsson 1980; 224).

5.4.2. Storage

In this section we are not so much concerned with the concept of item banks as a measurement system (Pollitt 1979;57) but rather with the problems of item storage in what might better be termed an "item pool" (Choppin 1978). We are thus not concerned with the design and constructiono of "the best possible test for any measurement situation" (Wright 1977;112), since this needs an explicit model to specify how a person and an item are supposed to interact, and this involves questions of 'ability' measurement as seen in the last section.

5.4.2.1. Objectives

Rather we are concerned with some of the objectives of item-banking as laid down by Wood and Skurnik (1969;8-9):

1. To build up libraries of first-class examination questions or items.

2. To familiarise more teachers with the ideas of modern examining, particularly the notion that an examination should be devised on the basis of a blueprint ,a document which codifies, in detail, the student attributes or behaviours which the examiner wishes to evaluate, and also the subject matter which is thought conducive to the achievement of these objectives.

3. To develop classifications of achievement which are universally applicable. Item banking was originally conceived as a method of monitoring attainment i.e. of achieving comparability of CSE grade standards, while allowing teachers to indulge in idiosyncratic teaching ideas and methods and examining practices. (Wood and Skurnik 1969; 76)

5.4.2.2. Creating the bank

Most importantly, the ideas and items in a bank should not be imposed by the tester or the testing service: "... the client, who is the teacher, should decide its contents ... each item in an examination should serve a stated purpose. It should measure some fragment of learning." (Wood and Skurnik 1969;13). Similarly, Buckley-Sharp and Harris (1970a and 1970b) and Buckley-Sharp (1973) emphasise the role of the consumer in the creation of banks; they feel that centralised item writing discourages individual thought by teachers, so they advocate individually owned 'accounts' in the item bank. Each individual makes deposits to his account and can seek 'loans' from other accounts provided he arranges to return a similar number of new questions at a later date (Buckley-Sharp and Harris 1970b).

Indeed, Buckley-Sharp's (1973) description of experience in this matter is revealing; he was concerned with the use of multiple choice questions in British medical schools, where the whole structure of objective assessment is determined by the small numbers of students in each school and by the great diversity of the courses. It had proved extremely difficulty to arrive at any working definition of the required aims or content of the medical undergraduate course, since this posed "vast organisational and philosophical problems". The resulting test material is therefore very extensive, and question banking was looked upon as a filing system for the data, if nothing else.

Importantly, and with few exceptions, questions are banked only after they have

already been used by examiners. This is in contrast to the other use of a question bank where questions are assembled, pre-tested and then used for definitive testing to the exclusion of new material. For Buckley-Sharp (1973) the variety of British courses, and the possible scope of medical assessment, precluded this "fossil" approach to assessment material in the question bank. Indeed, it is just such a "fossil" approach which encourages argument over the use of item banks. Hazlett (1973) has a similar system in Canadian medical schools.

Buckley-Sharp's (1973) selection procedure recognises only five categories (Hazlett has 57 variables which may be used to describe each item):

1. By subject mnemonic classification;
2. By department or examination reference;
3. By question analysis statistics;
4. By specified text profile; and
5. By specific question reference, which apparently is the most frequently used way of requesting items.

For the majority of cases the computer is used as a means of recovering neatly typed text, for material which has already been decided.

Newbould and Massey (1977) emphasise another practical use of item banks: that a multiple choice test item may only be cost-effective in the widest sense if it can be used on more than one occasion, partly because the development of item-writing skills is 'costly and time consuming'. Newbould and Massey's (1977) 'Computerized Item Banking System' is little more than an administrative convenience; they have a classical item analysis stage based on pre- and post-testing, and a 'specification grid' which sets out the desired balance to be held between the areas of knowledge to be sampled and the educational objectives set. So at the actual test construction stage, the user can see whether an item conforms to the appropriate specification within the grid, whether it is one of a group of items, whether it is a 'statistically sound' item, whether or not it has been used before, and by whom, and so on. It will be seen that such considerations are appropriate where large numbers of users wish to use a large centralised bank, but since the system still depends on classical item analysis, much work reamins to be done after an item has left the bank.

Lippey (1974) suggests that the main benefit of systems which bank items can occur when the items reside in a centralised collection shared by many users ,where

the overall quality of an item collection, properly modified as a result of experience, will continue to improve. This clearly views the item bank as a 'pool' in Choppin's (1978) sense and could be seen as an aid to the learning process (if items are administered often enough), if used to convey or reinforce ideas or stimulate the imagination (cf. Pollitt 1979;57).

Prosser (1974) also outlines a systematic approach to item production, which is, however, more concerned with the details of loading items into the item bank than with any serious theoretical considerations.

5.4.2.3. Item classification and selection

Of more interest is the question of item classification and selection. Gorth et al. (1971) make explicit the link between item banking and its "logical predecessor" objective banking. Often there is little distinction between the two, especially when the aims of item and objective banking are "to make teachers more familiar, in general, with modern notions of test construction including the classification of test items by categories which they measure e.g. behavioural objectives" (op.cit.; 245). For Gorth et al. (1971) objective and item banking require several operations including stocking of the bank, retrieving information from the bank and ,importantly, using the retrieved information in a variety of teting situations: "Using material from the bank consists of diagnostic testing within a course, placement testing, criterion-referenced testing within a course, pretesting for the different instructional treatments, or testing on a longitudinal basis using item sampling." (ib.;246)

Popham's (1978) Information Objectives Exchange (IOX) represents an attempt to make available to teachersinstructional objectives classified by grade-level, content, and taxonomical classification of objectives, though IOX is in reality more of an example of how to set about such a procedure than a fully-fledged objectives bank.

Walter's (1970) COMBAT (Computer-based test development center) has a bank of teacher-written test questions made available toclassroom teachers, classified in a similar way to IOX. Gorth's (1968) Comprehensive Achievement Monitoring (CAM) has developed a model of evaluation for curriculum evaluation and classroom management, which consists of longitudinal testing , using item sampling, of the specific behavioural objectives for a course. The CAM developers suggest that tests must be seen in a decision-making context, and that most decisions are classifiable into five types:

- Classification

- Summative evaluation
- Formative evaluation
- Instructional management
- Curriculum validation

(cf. O'Reilly et al. 1973; and Byrne 1976)

All the items in the CAM item bank are classified in at least three dimensions: their content, their taxonomic level, and the sequence in which they are taught in the typical school course. The latter point again reminds us that testing systems need to be related to some sequence of instruction. Odor (1985) suggests that one of the benefits of a complex classificationsystem is that while one may choose test items based on just one criterion, the end result is that if enough items are chosen, much extra diagnostic information can be deduced from an examination of pupil errors and item classifications.

However, the practical problems are many. Johnson and Maher (1984) describe the development of a search procedure (BROWSE) where it became apparent that important limitations existed, inthat such relatively simple search procedures retrieved only those questions which exactly satisfied the specified criteria. This is fine when the descriptive information on which a search is based can be precisely and unambiguously specified, but proved to be too limiting in mot cases. A more complex, but more versatile, thesaurus- linked queston banking system was developed in which superordinate terms (e.g. GRAPH) could retrieve all questions with more specific terms (e.g. PROJECTILE PATH, COOLING CURVE, DISTANCE-TIME GRAPH etc. in this case). The resulting flexibility seemed worth the extra effort.

Stodola (1974) suggests that to classify is to define major components of curriculum, since the classification categories directly reflect elements of the instructional program (p.67). He outlines the most commonly used attributes of questions for classification as : identification number, topic or content heading, index word, process, behavioural objective, and statistical characteristics. Of these, listing the major topics covered is said to be the most common.

Listing major topics is said to lend itself to a hierarchical approach in which one begins with major topics and subdivides these into subtopics and into further minor subtopics. This makes possible item selection by broad category or by detailed content specification according to the user's wishes, but does not provide for

definition of item content in terms of desired outcomes of instruction and does not set expected mastery levels.

Wood and Skurnik's (1969) classification of mathematics items in terms of knowledge, skills, comprehension, application and inventiveness represents a process approach (cf. Bloom et al.'s 1956 *Taxonomy*: knowledge, comprehension, application, analysis, synthesis, and evaluation) in which the intent is to arrange the categories from the simple to the complex and from the concrete to the abstract. Black and Dockrell (1980) adopt a similar, if not identical, viewpoint in their diagnostic framework (see chapter 4).

The advantage of such a system is that it helps to call attention to the need for questions measuring the so-called higher level reasoning processs. Problems arise, however, because of the dependence of item classification on a student's previous instruction. For example, the answer to a queston might require only knowledge if the student knew the answer, while for another who did not know it might require higher level reasoning. Moreover, any taxonomical system of this kind is not always clear as to how the classifications,which are listed according to a kind of psychological process, can be treated instructionally i.e. they are often too broad or too general.

So far as a behavioural objectives approach is concerned, there is usually less than complete agreement as to the level of specificity desirable in stating objectives. The principal difficulty for Stodola (1974;73) is that establishing a meaningful, integrated list of behavioural objectives covering many subjects is a huge task. For example, Toggenburger's (1973) Classroom Teacher Support System (CTSS) for high school history, comprising 3000 multiple choice items, is based on the fact that questions are merely one of any teacher's key instructional tools. The centralised bank of questions upon which CTSS is based is classified by content, difficulty level (three levels subjectively appraised), 'behaviour level' (knowledge or application of knowledge), keywords (up to three words or phrases per item), and a category termed 'X', which is used to indicate sources of materials, special formats, study aids and types of questions. The questions are constructed by history teachers and are strictly in the 'item pool' category i.e. an aid or stimulus to learning.

Simple classification and selection systems include Salisnjak's (1973) 'call-card' system and the computer-generated repeatable test system of Prosser (1973) and Jensen (1973).

Ansfield's (1973) Automated Examination Generator (AEG) and Libaw's (1973) MENTREX tutorial testing system are examples of fairly sophisticated classification and

selection systems. In MENTREX each person's test performance is analysed along several dimensions, such as subject content and cognitive skills, and a report is made to the individual of the degree to which he has mastered prescribed learning objectives. Individual tutorial feedback and references are provided to specific information sources to reveal what his test shows he has not yet learned, and tailor-made follow-up tests help the learner achieve and demonstrate an acceptable level of mastery of material not mastered on a previous test.

It will be seen yet again that to develop a sophisticated diagnostic testing system (the problems are similar to those of item banking) test items must relate to a course of instructin and a particular perception of the structure of the subject area . Stodola (1974;116) puts it thus: "No matter how good the system is according to some educational theory it will do the student no good if the classroom teacher rejects it because it is inconsistent with his own instructional approach."

The problems relating to item bank classification systes and categories of diagnostic testing are mutually dependent; any concern with more exact item classification is a step toward better definition of curriculum and permits better evaluation of the goals of instruction. Moreover, as Stodola (1974;118) suggests, development of detailed item classification systemsfurthers instruction aimed at highly specific objectives.

5.4.3. Operation and design

There are, according to Byrne (1976), two issues involved in the operating and designing of a bank of items:

1. What type of questions should the bank decide it will handle? The smaller the number of types, the greater the simplicity of design and operation; and

2. What format should the question take? Even for one type of question there may be many formats.

In the light of the second point, consider the Florida Agricultural Migrant Compensatory Reading Program (*Florida*) as outlined in Slocum (1972), Curtis (1972), Abbott (1972), and Wellens (1972). *Florida* has 55 basic structures or formats ('item types'). For example, the item-type:

"80-100 word passage + task + scoring guide - oral response"

was used for 104 questions measuring ten different objectives. From each basic

structure a number of prototypes was then developed (a prototype being an item type applied to a particular objective, of which there were 106, at a particular grade level). In the reading program 269 prototypes were developed which were then used as a basis for having questions written.

Even within the fairly strict limits set, however, problems can be seen: the example of item type given in Wellens (1972;31) is: "Components of critical thinking" passage + multiple choice item with four options; Major category:comprehension; subcategory:C. Critical reading & logic. Objective: the learner will be able to identify illogical thinking, inconsistencies, fallacies or discrepancies in a given selection.

However, the example given is:

"All the girls in my class live on Main Street or Broad Avenue. Most of the girls have older brothers. Which conclusion is false?

a.Some of the families on Broad Avenue have at least 2 children.

b.Some boys on Main Street have younger sisters.

c.Mary Ellen, who lives on River Road, is in my class.

d.Joan is in my class, so she must live on Broad Avenue or Main Street."

Here, the correct answer is said to be (c), but (a) and (b) could also be false; one has to be extremely careful in constructing items even within such an apparently strict framework.

So far as design is concerned, we would merely note that essay-type questions are just as bankable as any other, but Pollitt (1979) affirms that ,in the pursuit of reliable examining, all item banks so far constructed have limited themselves to multiple choice or other thoroughly objective single-mark items. Though Willmott (1976) adds that objectivity of measurement and development of a flexible measuring instrument has been seen to be "the key to the inner working of the item bank."(p.42)

5.4.3.1. Item banking in L2 learning

Popyuk's (1980) development of a system of item banking at the Language Standards Control Detachment (LSCD) of the Canadian Forces Training System headquarters was based on the assumption that "language test items are inherently stable and item characteristics will remain relatively constant provided pertinent variables are controlled" (p.47)

Language tests were needed to measure the language proficiency of military personnel in their second language, and batteries of tests were created to measure the "four integrated language skills". Language proficiency levels were established from level 0 (no measurable proficiency) to level 5 (native-like proficiency), which were said to be "carefully defined" for each skill and became the external criteria against which all tests were to be normed.

According to Popyuk (1980;48) the objectives in developing an item-banking system were:

1. To build up a library/bank of test items at different levels of proficiency, employing various testing techniques;
2. To obtain data on individual test items which would have 'universal significance';
3. To permit rapid assembling of pre-normed tests of items selected from the bank.
4. To analyse individual item performance as directly related to the external criteria; and
5. To allow for interchangeability of items from one test to another.

To this end, item bank cards were created with linguistic and statistical information on them. The statistical information was based on classical item analysis techniques, continually updated, so that it is difficult to see how the objectives of achieving 'universal significance' and obtaining pre-normed tests could have been satisfactorily achieved.

Moreover, an investigation of the pattern of omissions in the tests revealed that the position of items in a test is a more significant factor with respect to omissions than is item difficulty - it was found that more candidates attempt to answer more difficult items positioned at the beginning of a test than easier ones situated further in the test. One implication of this could be that a candidate's proficiency level, on a test with a wide enough spread of difficulty in the items, could be determined simply on the number of questions unanswered, regardless of whether they were right or wrong

Popyuk (1980) seems to expect too much of his bank; an 'international item bank' is being accumulated from sample populations drawn from NATO representatives. Should such a bank ever come to fruition it will, however, only be strictly applicable to the NATO population from which it was drawn, and will tell us little about L2

language proficiency for other groups.

5.5. Assumptions and Objections

Apart from the problems of simply gathering an adequate number of items and cataloguing them in a satisfactory manner, a major requirement of item banks is that there be "a theory ennabling the meaning of all scores on any test which may be constructed from the bank items to be defined in advance from the obtained item data." (Pollitt 1979;57)

Moreover, the potential of item banking can only be realised under the assumption that a test is required to measure accurately a single trait. 'Latent trait' models of test performance assume that there exists a single scale on which all items can be placed, ordered according to their difficulty, and that this order is fixed for all persons attempting the items. Furthermore, all the examinees can be placed on this scale, ordered according to their ability, and this order is independent of the actual items attempted. There is thus a single scale measuring the latent trait which measures both difficulty and ability simulataneously. (cf. Pollitt 1979; 58)

5.5.1. Latent traits: pros and cons

Because item banks require 'object-free instrument calibration and instrument-free object measurement' (Wright 1968;87) in order to generalise measurement beyond the particular instrument used (i.e. the calibration of test easiness must be independent of the particular persons used for the calibration, and the measurement of person ability must be independent of the particular test items used for measurement) much of the theoretical discussion of item banks centres on the statistical models underlying them.

The claimed advantage of latent trait models as a basis for constructing item banks include, for example:

- the fact that they would allow different institutions toconstruct different tests yet still award equivalent grades (Pollitt 1979;64);
- the fact that 'absolute' academic standards can be compared over time (Goldstein 1979;217);
- above all, perhaps, that an individual's 'true ability' can be measured.

All tests made up from such a bank are automatically equated; since a person's score on any test can be converted into an ability estimate on the common bank scale, any

group of people can be given a test made up of items particularly suitable for them, yet all the results can be compared with each other.

Baker(1974) shows how classical item statistics can be misused in a CATC environment: creating tests and administering them to heterogeneous groups of pupils severely strains the classical item improvement cycle. the strain arises from the natural inclination to pool the item data both across tests and across groups and to keep an agglomerated set of statistics for an item. The propensity to pool data in this fashion is a manifestation of the assumption that items are invariant.

The particular statistical procedures involved in item banking (especially the use of probabilistic test models such as that of Rasch 1960) will be discussed more fully in chapter 5 (see,for example, Rasch 1960,Wright 1968, Baker 1974,Wright 1977, and Pollitt 1979).

Unfortunately, there are a number of objections to the use of latent trait models. Advocates of some sort of latent trait model tend to point out the difficulties involved in two-,three-, or n-parameter models (cf. Osterlein 1983), or else emphasise that a Rasch model, for example, will not work if

- the test is highly speeded, because the model assumes that everyone gets a fair chance at every item;
- the test is essentially one of knowledge (rather than 'ability') - there is no dimension on which to place the items uniquely; and
- the test confounds two or more poorly correlated dimensions i.e. it measures more than a single trait. (Pollitt 1979;60)

Baker (1974;151) points out that many CATC systems assume that the item pool is homogeneous and that the measuring capabilities of all subsets of the item pool are equivalent; this may not reflect the reality. Moreover, the Rasch model assumes that all items have the same discriminating power. (Baker 1974;156)

Hambleton and Cook (1977) discuss various aspects of latent trait models and suggest that, as with all test models based on restrictive (i.e. strong) assumptions, practical applications are limited "because of the failure of mental test data to satisfy the assumptions underlying the various test models. In particular, the assumption that all item discrimination parameters are equal is restrictive, and substantial evidence is available which suggests that unless test items are specifically chosen to have this characteristic, the assumption will be violated (Hambleton and Cook 1977;83). Wood and Skurnik (1969), the British pioneers in item banking, allude also to the 'spurious

sophistication' involved in any obsession with statistical models, and go on to state that items chosen solely on their statistical characteristics may result in a reliable test but not necesssarily a valid one, "and it is validity that matters most" (p.60)

However, there are wider objections to the whole question of using latent trait models in general and item banks in particular, as discussed most notably by Goldstein (1979). The following points can be made: firstly, there is a problem (which occurs also in discussions of factor analysis) in speaking of 'ability', since this merely gives a label tosomething which is in fact defined in terms of statistics (cf. Wright 1977;103). Comparing the Rasch model with factor analysis, Goldstein (1979;212–213) emphasises the fact that a strong objective reality cannot be attributed to factors derived from such techniques, where, in the Rasch model for example, a set of indicators (responses to individual items) is related to a set of unobservable factors or traits ('ability').

Secondly, the Rasch model, according to Goldstein, assumes (a) that the relative difficulty of the items in a test is the same for all individuals i.e. we would require that, despite different experiences, learning experiences, learning sequences etc., the difficulty order of items was the same for every individual; and (b) that there is 'local independence' i.e. for any individual, the response to an item is completely independent of his or her response to any other item.

Thirdly, there is the 'philosophical' issue of whether the model should fit the items, or the items should fit the model. There is a danger that a test will be defined in order that it fit the (Rasch) model, without any consideration of whether the model itself might not possibly apply. Hence, says Goldstein (p.216),the attainment which is being assessed is effectively defined by those items which happen to conform to the Rasch model, with no guarantee that the result will have "a real-life interpretation".

Fourthly, even if we could construct a well-fitting Rasch model in a situation where we might expect this to be possible, there is no necessary educational reason for preferring it over any other method of test construction. It seems *a priori* unlikely, for example, that a reasonable and fair set of itemscan be found which appear in the same difficulty order for all testees. Indeed the essence of many educational systems is the diversity of approaches whose actual aim is to create differential attainments among otherwise similar pupils, for example by way of the ordering or teaching or as a result of different pedagogical practices.

Finally, there is an "inherent flaw" in the item bank concept which, according to Goldstein, would make it unworkable in practice: if we suppose that each of the items

in a bank has a prescribed difficulty value, then it is strictly meaningless within the context of the Rasch model to speak of one item as being *more applicable* to one point in time rather than another. The only meaning which can be attached to such a statement must be in terms of difficulty value, thus an item bank which is designed so that out-of-date items can be replaced is a strictly non-Raschian concept. Similar logic applies to tailored testing procedures, where it is claimed that items can be selected from an item bank to suit different curricula. A further consequence of this would be that there is therefore no absolute basis on which comparisons can be made over time. (Goldstein 1979;217-218)

These are serious objections indeed, but we should be careful to distinguish between claims based primarily on statistical considerations and claims based on what might be called 'content' considerations.

Indeed, much of the debate over latent trait models and item banks seems to involve an approach to the question based firmly on one or other of these two emphases. Wood and Skurnik (1969) suggest because any test or examination can be considered to bea sample of items from the universe of all possible items which are eligible to be included in the test, the question is one of how much we can legitimately infer about an individual's global capacity from his performance on any one attainment test.

On the one hand we have the notion of a universe-defined test (see chapter 4) and on the other hand we have a reliance on 'latent traits' representing a hypothetical dimension of knowledge or skill.

The problem is that so long as what the test is supposed to be a measure of is conceived to be an ideal quantity, unmeasurable directly and hence undefinable operationally, test validation will be a troublesome task.

Wood and Skurnik (1969) claim that these two opposite points of view are reconcilable "providing an attempt is made to describe what it means in terms of specific achievement." This, however, would seem implicitly to reject the notion of latent traits and return to the world of operational definitions.

Baker (1974) offers two practical suggestions:

1. CATC systems should calculate item statistics only when the tests generated deal with a limited content area, the items are not used for sharply different purposes within that content area, and the population of pupils tested is consistent across multiple uses of

the same item.

2. On the issue of equivalent test forms, in order to treat the assembled tests as randomly parallel, it is necessary that the item pool be a well- defined universe of items, not just a collection of items. Simply selecting items at random from a collection of items does not ensure the creation of randomly parallel tests. The implication for CATC is that there should be a relatively large number of items for each combination of content descriptors. Thus when items are selected at random they are obtained from a well-defined population of items. The obvious trade-off is then between the number of descriptors employed and the size of the item pool.

In spite of the difficulties involved, however, it would seem that the concept of item banking is a useful one provided one realises its limitations i.e. it is *not* going to provide some magic index of ability, but it is a useful source from which to build tests, especially if related to a particular course of instruction.

The formation of an item bank depends first and foremost on the availability of trustworthy item parameter estimates (Wood 1976; 252). Once the means of obtaining such estimates are available the next step is to calibrate items on a common scale; this calibration procedure should be applicable in different circumstances so that items given in one administration of a test can be compared with different items given in a different administration of a test and so that the two sets of items can be calibrated on a common scale. The procedures for carrying out such calibrations are relatively straightforward, though it is important to be systematic. McBride and Weiss (1974) deal specifically with the calibration and updating of an item bank and envisage seven basic steps:

1. Define the population of interest

2. Compile an initial development test and norm it on a representative sample from the population

3. On the basis of the norming data, construct a calibration/criterion subtest

4. Construct several long secondary development tests incorporating the calibration/criterion subtest

5. Norm these development tests on large representative samples from the population

6. Perform item analysis of the development tests, employing calibration/ criterion subtest scores as the criterion

7. Select items for the item pool on the basis of the item analysis

data

Clearly the difficulty here is with the amount of work involved rather than with any controversy as to procedure. The main point to note is that great care is needed in the construction of the anchor test. McBride and Weiss think the test might need to be as long as 40–60 items, which would in normal circumstances place quite a burden on the groups who would have to take the anchor test in addition to other tests. A further demand which is not easy to meet (and which is also a feature of classical test development) is that the norming tests need to be administered to "very large samples" from the population in order to achieve stability of the item parameters. (On the issue of *internal* versus *external* anchor tests, see the discussion above under Test Equating.)

Haksar (1983; 262) outlines a procedure for developing a Rasch calibrated bank for criterion-referenced testing. She suggests the following steps:

1. Identify criteria and place them on a difficulty/ability scale
2. Set an appropriate test
3. Calculate the score-measure conversion table for this test
4. Mark the criterion level together with the error limits on this table

One can then take decisions on criterion testing at whatever level of certainty is required by setting the error limits to the appropriate multiple of the standard error. The problem here is that the full calibration procedure is presupposed, and the difficulties that remain are the difficulties associated with any criterion-referenced test, especially the decision as to what constitutes mastery/non-mastery.

Technical descriptions of linking items to form a bank, without necessarily claiming totally invariant estimation of parameters, can be found in Woods and Baker (1985; 128–131). This is a fairly simple arithmetic procedure by which differences in item difficulty estimates are averaged out across different administrations of a test. Though this is the least controversial way of developing an item bank, it is a procedure which has moved on from the stronger claims of Wright and Stone (1979), for example, where it seems to be implied that continuous evaluation of parameter estimates is not necessary. This, however, appears to be the current trend in thinking on item banks; it is a procedure supported by, for example, Theunissen (1987), who also introduces the rather complex notion of 'test information' as a test design tool. What this appears to amount to is the use of a combined ability and difficulty function

to help assess the optimum composition of a test. This, however, depends on reliable parameter estimation (as always) and in practice looks little different from the already established principle that the optimum test for an individual will consist of items at the same level of difficulty as that individual's ability. It is difficult to see the virtues of this procedure.

The essence of Theunissen's (1987) approach is the establishment of the smallest number of test items from a bank which will satisfy a specified criterion. The core problem is still that of establishing mastery/non-mastery points, an issue which has been addressed by Hambleton and Swaminathan (1985; 260-262). Hambleton and Swaminathan make clear what Theunissen does not, namely that careful calibration procedures and evaluation of fit are as important here as in any other area of item banking. Moreover, a criticism of this approach (and indeed of adaptive testing generally) would be that tests resulting from such an 'optimal item selection' strategy may lack content validity because items are selected on the basis of their statistical characteristics. It requires great confidence in our interpretation of the 'ability' parameter to be able to ignore this criticism entirely.

The most detailed description of item bank building using link items is found in Wright and Stone (1979; 98-106). Essentially this is an arithmetic procedure which involves using certain core/link items in different forms of a test to rescale the difficulty estimates obtained. The complexities of the procedure arise from the number of links being used, though it is not clear in Wright and Stone's account how Wood's (1976) criticism that *slippage* is unavoidable each time a recalibration takes place has been avoided.

5.6. Conclusion

There are good grounds for adopting the one-parameter Rasch model for language test data. However, the final justification for the use of any model must lie in its conformity to the actual behaviour of test items. Whatever purpose the item bank will serve, we must have empirical evidence that the model is functioning in a way that reflects our intuitions and insights into the nature of the test content.

The next two chapters attempt to put into practice the principles outlined so far. Chapter 6 uses the insights of Chapters 2 - 4 to establish a test with prior content validity, while Chapter 7 uses the concepts discussed in this chapter (Chapter 5) to evaluate the item bank concept on the basis of actual test performance.

CHAPTER 6

METHOD OF DESIGN AND CONSTRUCTION OF TEST ITEMS AND ITEM BANK FOR PLACEMENT TEST PURPOSES

6.1. Introduction

6.1.1. Summary

The construction of an item bank to test reading in English as a foreign language offers the possibility of flexible approaches to testing as well as the hope of more 'objective' measurement than has been possible in the past. There are, however, a number of problems associated with the construction of such a bank. The first set of problems is related to *content:* can we define reading in such a way as to ensure unidimensionality, or at least to be confident of knowing how many dimensions we have? The number of dimensions will then be reflected in the number of separate banks of items we need to construct. In Chapter 2 it was suggested that it is possible to views reading in English as a foreign language as a single dimension (or construct) which embraces a wide range of activities, from using grammar in tightly controlled contexts to more global reading comprehension activities.

The construction of test items also raises the problems of validity which were addressed in Chapters 3 and 4. *A priori* validity is an essential component of any test construction activity but is particularly important in the item banking context where much use is made of such concepts as item *difficulty* and person *ability.* Although these concepts have strictly statistical meaning they are nevertheless related to common sense interpretations of those terms ; in other words they represent *semantic interpretations* of *syntactic definitions* (Lord and Novick 1968;17). In Chapters 3 and 4 various ways of ensuring validity in approaching the testing of ability were suggested.

The second set of problems relates to more technical matters in connection with practical use (e.g. the labelling of items) and with the limitations inherent in traditional test statistics. It was concluded in Chapter 5 that latent trait models of testing promise to overcome many of the difficulties associated with traditional test statistics.

The conclusion from our analysis so far is that even if we do not know, perhaps can never know, exactly what 'reading' might be in its psychological sense, we can at least create operational definitions of reading which allow us to procede on the basis that the dimensions of reading are known with sufficient confidence. Furthermore, by emphasising the content validity of our tests at the test construction stage we can

avoid many of the problems traditionally associated with claims to test people's 'ability' in a particular subject. In addition, we can attempt to control the difficulty of the test items not only by reference to content but also by rigorous attention to item construction. The use of item response theory should overcome the inherent technical problems in creating a bank.

6.1.2. Immediate objectives

The questions we should now ask are therefore:

1. Can test items be written to accurate enough specifications that we can establish dimensionality before statistical analysis?
2. Can a workable bank of items be constructed, by whatever method?
3. Does item response theory lead to the development of flexible testing?
4. Does an item response model possess clear advantages over traditional models for this particular domain?
5. On the basis of test results, how many dimensions are there to reading in English as a foreign language i.e. how many scales do we need to account for the data?

As far as point (4) is concerned, 'advantages' will be taken to mean increased stability of test statistics, and not to the construction of an actual bank, since the latter is not a situation which classical statistics are designed to cope with. For that reason no comparison was made between classical test statistics and Rasch statistics for item bank construction.

6.2. Background

This Chapter describes the construction of a small item bank arising from a practical testing situation. The present author, acting as a research assistant on a British Council funded project in association with Dr A. Davies of Edinburgh University, visited the Universiti Sains Malaysia (USM) in October 1986 to assist in the development of a placement test for matriculating students at the university.

About 1500 students matriculate each year, and although USM is a 'science' university, about half of the students follow courses in other faculties (arts, languages, history, law, religion, business and architecture being the main ones). The need for a placement test arose because all USM students would, from July 1988 onwards, be

required to demonstrate a minimum competence in English before being allowed to graduate in their chosen discipline. Accordingly, all matriculating students would have to be placed at one of five levels to determine their English tuition requirements:

Level 1:	Elementary
Level 2:	Low intermediate
Level 3:	High intermediate
Level 4:	Advanced
Level 5:	Exempt

In the English language programme as described by USM prospectuses, Level 1 aims to "focus on elementary reading comprehension", Level 2 to "focus on comprehension of scientific/various texts", Level 3 to "focus on comprehension of university specialised texts," and Level 4 to "focus (60%) on oral skills." Two immediate points should be highlighted here: firstly, that the requirement is essentially for a *reading* test and secondly, that whatever placement procedure is decided upon must cover a very wide range of ability. It should also be noted that for the current study Level 4 could be treated as the "exempt" level since a pass at Level 3 is all that is required for graduation purposes, though Level 4 course must be taken as a university requirement if the student is not exempt at Level 5. Clearly, the cost of making a mistake in placement between Levels 4 and 5 is nowhere near as great as the cost of making mistakes at lower levels.

In essence, then, we are concerned with the development of a test of reading. The final version of the test does include a summary-writing section and provision for oral assessment through an interview for candidates at the top of the range, but these sections of the test will not be discussed in any detail here.

Matriculating students in the years preceding 1988 have all been placed on the English language teaching programme, so why was there a need to change existing procedures? There are two main reasons for this: firstly, it was generally felt that misplacement was occurring on too large a scale for comfort (although anecdotal evidence suggests that this may have been exaggerated). Students were placed using a variety of measures, mainly school leaving exams which were, in some cases, taken years before matriculation and therefore no longer a true reflection of a student's true English level. In any event, there are two school leaving exams (SPM 322 and SPM 1119) which not all students take, so no single satisfactory test was being administered to all matriculating students. A new test, it was planned, would

introduce uniformity to the procedure and minimise the risk of assigning mastery to non-masters and non-mastery to masters, a problem felt to have been exacerbated by previous placement procedures.

The second, and ultimately more significant, reason for the development of a single placement test was that such a test should, once piloted, be available for use in other, similar, situations in different universities not only in Malaysia but elsewhere in South-East Asia and perhaps further afield. This is precisely the sort of situation that item banking in general and item response theory in particular are designed to handle - different populations taking the same test. An interesting question arises here: how different does a population have to be before item analysis statistics break down? This question is taken up again in Chapter 8.

The programme laid down for the development of this test required (i) a pilot version of the test; (ii) a revised pilot version and (iii) a final version. This reflects normal procedure; where one would expect IRT to be of advantage in a situation like this would be in eliminating the need for large-scale trialling and re-trialling: a number of items could be calibrated as 'links' in a chain (Wright and Stone 1979; 98). The present study therefore takes advantage of this situation by following the traditional procedure and using it as a check on the IRT procedure. It will be remembered from Chapter 1 that one of the features distinguishing an item bank from an item pool is that in the former, but not in the latter, items will have an individually calibrated 'difficulty' which will remain stable. However, if the population tested is large enough and homogeneous enough, then an item pool can become an item bank (in a limited sense) using traditional statistics. The need for sample-free statistics only becomes apparent when generalising from small, possibly atypical, samples to larger populations. Sufficient numbers are involved in this study (between 1100 and 1300) to allow us to treat the group as the population and therefore to accept the traditional statistics as the norm - IRT statistics must be at least as acceptable if they are to be given consideration. When it comes to extension to other populations this assumption with respect to the relationship between sample and population will not hold. If however, IRT can be shown to work in this situation, then we are justified in extending its use.

6.3. Test design

6.3.1. Situational constraints

One or two factors out of the test constructor's control pose special constraints on the construction of this test. These constraints are, however, not always negative. For example, as mentioned above, the requirement is for a reading test (in essence): "The critical level of ability is that to be attained at the end of the Level 3 course with its primary focus on an ability to understand scientific/specialist textbooks and with its secondary focus on ability to write notes and summaries." (Moller 1985; 5). Further requirements of the test include the important provision that "a sound grasp of basic English grammar is necessary before Level 3 work can be effectively undertaken" (ib.). Moreover "part of the test content must relate *back* to the students' previous English language learning experience and part must look *forward* to the criterion skills expected of students at Level 3 (ib.). As mentioned earlier, listening and speaking are not important in this context, at least not until the critical level (Level 3) of reading ability has been achieved.

The test needs to be easily administered, since it will be given to up to 1600 students at one sitting, prefereably easy and quick to mark, and, of necessity, secure. Clearly a computer-markable multiple-choice format must be the first choice.

Further provisional constraints on the content of the test are described in greater detail in Moller (op. cit.; 6) and stem largely from the fact that the test will be in three parts:

1. Part 1: a reading test, testing knowledge of basic grammar. This should probably be pitched at the level of the OUP Placement Test (Part B) or the British Council Mini-Platform Writing Test. Maximum time 1 hour.

2. Part 2: a reading test, testing comprehension of scientific and specialist texts in accordance with skills laid down in the syllabus for Level 3. Only students who have been exempted from Levels 1 and 2 on the basis of Part 1 need take or be assessed on P{art 2. There may be two versions of Part 2 - one for science, medical, building and engineering students and one for arts, mass communications and management students. There could be some texts/sections common to both versions. Maximum time 90 minutes.

3. Part 3: a test of oral interaction with a teacher.

6.3.2. One test or many?

It is possible to make out a case that there should be two versions of a test – one for 'arts' students and one for 'science' students. This, after all, was the basis for the British Council ELTS test in its first incarnation. The argument would be that different groups have different needs. This of course is unobjectionable in itself; the problem, however, is that it breaks down in practice (see e.g. Criper and Davies 1986; 115). How, for example, once the principle of separate needs has been accepted, could one reasonably justify medicine and engineering students taking the same test because they happen to be science students? And if a separate test is then developed for each why should surgeons take the same test as obstetricians? Analagously, why should thoracic surgeons take the same test as neurosurgeons? The argument quickly becomes absurd, but no satisfactory resolution of the dilemma is available. In practical terms it makes sense to try where possible to reduce the number of versions of a test which are available (not of course the components of the test); in the current context, specialised training has not begun, thus any narrowing of minds which EAP testing is designed to encourage will presumably not have taken place. Moreover, in view of the fact that both the arts and science English programmes often follow similar lines , as will be seen in the next section, it would be difficult to say that one group would be disadvantaged by having a test versiion which was neutral in content, or indeed that any special advantage would arise from being tested on texts in one's own discipline. The principle of separate versions is attractive at first blush, but experience, particularly with ELTS, suggests that it is a principle less easy than it might appear to defend both in principle and practice.

After discussion with those involved, the decision was accordingly made to keep to one version of the test for all students. Later extensions, especially if IRT proves of value, could be added if it was thought necessary.

6.3.3. Test structure and the placement problem

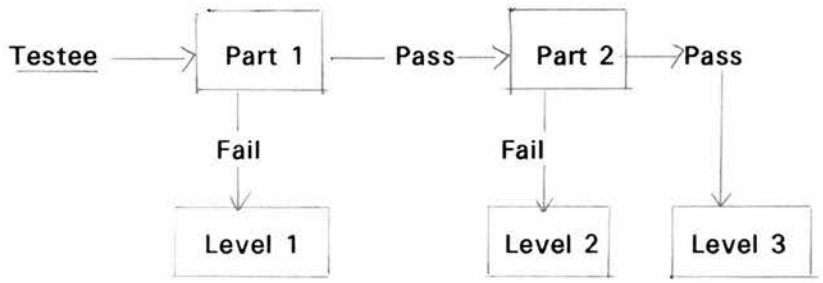
In discussion of the test from this point on we shall try to limit ourselves to the reading part of the test. The summary-writing part of the test (Part 4) and the oral interview will not be dealt with in any detail. Reference to the 'test' will mean the first 3 Parts (i.e. the 'reading' parts) of the test.

The local constraints mentioned above suggest, broadly speaking, a two-part test to reflect the perceived dimensions of reading in this situation: "basic grammar" reading and "reading skills". Initially these were entirely defined by the syllabus. From the discussion in Chapter 2 it will be remembered that attempts to separate reading

into anything other than 'low' and 'high' level dimensions (on the psychometric side) or to identify the components of comprehension in any way that allows for trainable elements to be isolated (on the psycholinguistic side) are dubious to say the least. The two-part test provisionally envisaged here would thus reflect quite neatly the two broad levels identified by Lunzer *et al.* (1979) and Hillocks and Ludlow (1984). The argument would further be that, because these represent low and high level skills, and because, in the project proposer's own words, the first part should be a "reading test, testing knowledge of basic grammar" (Moller 1985; 6) then an automatic selection process will take place, thus doing at least part of the job of placement. In other words, students failing Part 1 will be assigned to Level 1 English; those passing Part 1 but failing Part 2 will be assigned to Level 2 English; those passing Part 2 will be assigned to Level 3.

The problems here relate to exactly what it is that is involved in the placement procedure (we shall return to the content of the test shortly). Is placement to be seen as a series of hurdles which the candidate has to clear? This is the simplest model and may be represented diagrammatically thus:

Figure 3
Simple placement model

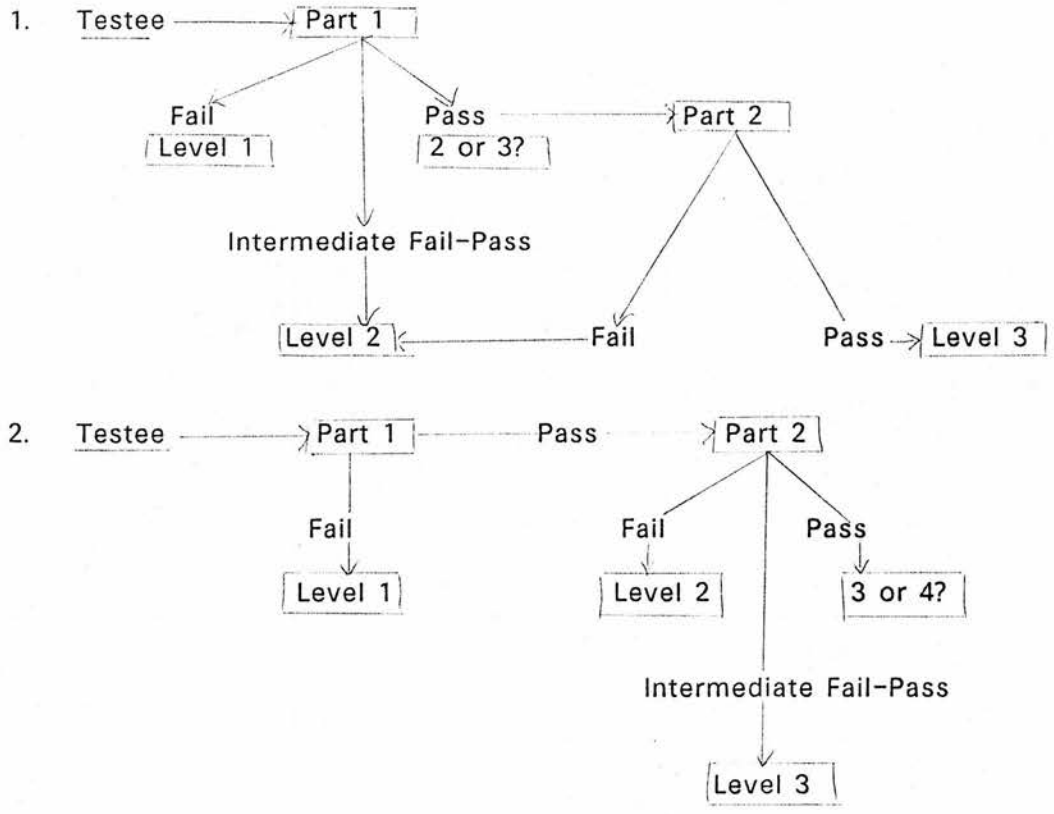


Exactly what is meant by "Pass" and "Fail" here would seem to be an empirical question, unanswerable *a priori*. The major difficulty seems to be that on closer inspection either Part 1 or Part 2 will have to shoulder a disproportionate amount of responsibility for selection. In diagrammatic form again, the reality has to be one of the following two situations:

Figure 4
Alternative placement model

cont.

Figure 4
Alternative placement model



The situation is confused because of the grey area between Parts 1 and 2. Can there ever be strict hierarchies of difficulty or learning for situations of this kind? Horne (1984) suggests that this is unlikely (see the discussion of this in Chapter 2 above). In other words the placement problem poses a particularly thorny question which should, in theory, be answerable by IRT, which *requires* a strict order of difficulty for test items. Traditionally of course scores from parts of tests are merged to give one score and no attempt is made to say what this might mean - in the current context no attempt would be made to say which hurdle the testee has managed to jump, save in very general terms. The promise of criterion- referenced and diagnostic testing (see Chapter 3) is that this problem is solvable, given suitable test construction and test analysis. Again, the ELTS experience suggests that practical problems confound our good intentions (Criper and Davies 1986), though it should be said that in spite of apparent specificity of design, there was little attempt with ELTS to follow the implications of its own statements about test design in that specifications for test items were often too narrow or too broad to be of any use (see discussion of content validity in Criper and Davies op. cit.).

Traditionally, then, the placement procedure has been a matter of deciding on cutting scores, using the overall mark (usually) from perhaps one or a series of tests.

In this way placement says little about what indiduals can or cannot do and places a lot of faith in what we might call 'superordinate' abilities - e.g. by saying that such-and-such a test is a test of 'language ability'. This is inevitable if one measure and one score are used. Thus, using only traditional methods it would not seriously matter if, say, an individual failed Part 1 but passed Part 2 (assuming the hierarchical relationship that 2 is more difficult that 1 in some sense) because it is the overall score that counts. True, in this case the validity and probably the reliability of the test will be compromised; but this is the strength of traditional analysis on norm-referenced lines, namely that it tends to swallow up small mistakes in the interests of a larger view. The criterion-referenced problem - and the placement problem here - is that we may expect too much of one test to tell us what is going on. Hence the emphasis, as argued in Chapter 3, for a thorough analysis of test content.

It might be argued that in this situation we should not worry too much about false negatives on Part 1 - after all, we can still put the results of both parts together and use the overall test score, as in the traditional analysis, to guide our placement decision. There are two counter-arguments here. First, it is true that scores can be combined, but if we think that we are testing separate things in the two parts then we are logically required to keep the scores separate. This of course almost never happens in practice, but it is still logically wrong and therefore not defensible. Incidentally, reliability and validity will both be compromised, as mentioned above.

Second, and more important perhaps in the present context, placement is actually serving a function here in relation to a syllabus - it was stated that "a sound grasp of English grammar is necessary before Level 3 work can be effectively undertaken" (Moller loc. cit.). In other words if we fail to weed out those who will not be able to cope with the work at the level at which we place them, then we are failing in our duty to them. In a nutshell, the placement test should perform a very definite diagnostic role.

The main problem associated with placement is that already alluded to of assigning mastery to non-masters and non-mastery to masters. Diagrammatically we wish to make the number of candidates falling in the top right and lower left quadrants as small as possible.

Figure 5
Ideal outcomes of mastery decisions /cont.

Figure 5
Ideal outcomes of mastery decisions

	Master	Non-master
Master	✓	✗
Non-master	✗	✓

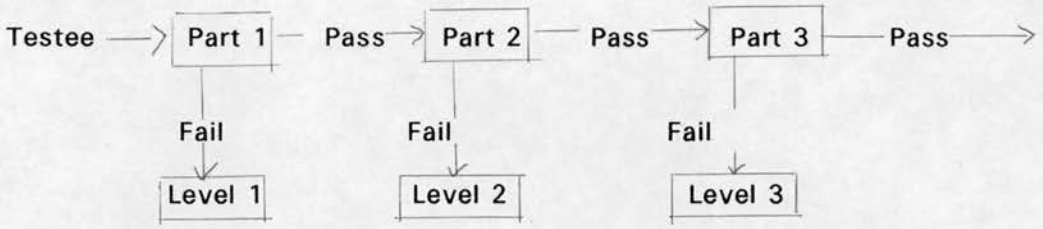
(✓ = desired outcome; ✗ =undesired outcome)

In reality of course we cannot eliminate wrong decisions entirely but we *can* decide on which side of caution we wish to err - should we make it more easy for students to obtain a pass, thus increasing the numbers of false positives? Or should we make it more difficult to obtain a pass, thereby increasing the number of false negatives? Our answer depends largely on the consequences of our wrong decision-making. In this particular case one might argue that we should err on the side of leniency, since we want students to rush through their English course and be able to spend more time and effort on their chosen disciplines. This view would be persuasive were it not for the fact that a pass in English at Level 3 is presumably required not just as a bureaucratic requirement, but as a genuine attempt to raise standards - otherwise why should the placement test have assumed the importance it evidently has? An overly "humanistic" approach may be counterproductive anyway, in that if false masters are assigned to the intermediate levels (2 and 3) then presumably they will find it difficult to cope with the work at those levels and thus find it more difficult to obtain the much-desired pass. On the other hand, what are the costs of making a wrong decision in the other direction i.e. of making the test more difficult and increasing the chances of assigning non-mastery to masters? Well, a certain amount of frustration and boredom perhaps; but false negatives should be able to move more quickly through the levels.

No decision as to pass/fail levels need of course be final - at least not if we err towards a more thorough-going norm-referenced approach. However, it is incumbent upon us to attempt to provide a test which is not only meaningfully interpretable but which also does the job it is meant to do. The critical area in this situation seems to be Level 3; we should thus aim to be confident that those who pass the Level 3 part of the test have a strong likelihood of being true masters.

In summary, then, the ideal placement test battery for a situation such as that described here would look something like this:

Figure 6
The placement model and the test battery



The test will thus be constructed on the basis that Part 1 represents typical work at Level 1 and a pass is sufficient indication that the groundwork has been covered to ensure a reasonable chance of success with Level 2 work. Similarly a fail means that (remedial) work at this level needs to be done. A similar argument applies to the other two parts, each being designed to reflect the work covered at a particular level.

Objections to this type of design are largely practical: the problems of wrong decision-making as mentioned earlier will be one source of difficulty. Another problem will be that strict hierarchies may not prevail. Yet another problem is that, in all likelihood, the scores from the separate parts of the test will be pooled and the effort will have been in vain. Nevertheless, the principle would appear to be sound: if we are confident of the validity of our different levels, we should be confident of our ability to differentiate testees on some meaningful scale.

One other practical point needs to be made: if the placement test really does work as intended and as outlined in the above diagram, then it would not be necessary for all matriculating students to take all of the test. There are two possibilities: everyone takes Part 1; only those who pass Part 1 need take Part 2; and only those who pass Part 2 take Part 3. THere are two objections to this: firstly, the administrative burden may be increased in this way, since there would have to be tests on at least three different days (to allow time for marking). Secondly, and more importantly, there is likely to be a psychological block for those students who found the test easy but for some reason failed; in other words, provision would have to be made for re-assessment or for allowing those who felt able to take Part 2 (etc.) anyway. In which case, why separate them in the first place? In connection with this we should also add that we would be placing very great confidence in our test-writing abilities if we proceded in this step-wise fashion (though this does not mean that our aim should still not be to attempt to construct the test in this fashion).

The other possibility will depend on the success of IRT in the test analysis: all the items can be entered on computer as a bank and the testees need take only those items at their particular level of ability. This principle has been discussed in Chapter 5 . Meanwhile the obvious point should be made that this would only be possible if enough computer terminals/time could be made available.

In conclusion, as a matter of principle the test should attempt to reflect the three levels to which it is meant to assign students (as argued in Chapter 3). Whether these constitute in reality three separate dimensions (in the sense discussed in Chapters 1 and 2) or even just two dimensions or perhaps only one will need to be investigated by analysis after the event. Initially we expect, in the light of conclusions drawn in Chapter 2, that there will be two dimensions at the most (represented by Parts 1 and 3). If it transpires that Part 2 forms an easily recognisable separate dimension, then we shall be one step nearer producing a hierarchy for reading in English as as foreign language. However our initial expectation, for reasons discussed in Chapters 2 and 4, must be that Part 2 will show more characteristics either of Part 1 or of Part 3. Furthermore, it is expected that the data will be describable in terms of a single dimension - though whether we would actually want so to describe it is largely an educational rather than a statistical question .

6.4. Design for pilot tests

Testing tradition follows a pattern which generally demands at least two runs of a test to establish its final form. This has usually been for the purpose of eliminating items which function poorly on the basis of facility value or, more usually, discrimination index. The present test is no different in this respect; there seems to be little sense in assuming the correctness of IRT and drastically reducing the numbers of items tested or the occasions on which they are tested. Indeed, it is part of the purpose of the present study to see if such a method would be workable. In broad outline, then, the current test procedes on the basis that each trial of the test acts as a sieve or screen; accordingly, larger numbers of items have to be written than are actually required. This is normally a disadvantage. In this case, however, the extra items will serve as a useful check on the validity of the procedures. After all, provided an item meets the requirement of unidimensionality (i.e. it is not about a completely unrelated topic) there is a sense in which there can not be a "bad" ITR item - merely a very "difficult" one (or a very "easy" one). This is implicit in traditional analysis since the basis for rejection of an item is related to "facility"; however, items would normally be labelled "bad" and a reason found for their performing badly.

To maintain an even balance, and to keep the length of the test manageable, Parts 1 and 2 in the initial version comprised 50 multiple choice items each, while Part 3 comprised 40 multiple choice items. Two parallel versions were designed for piloting purposess, thus giving a total of 280 items. From these 280 it was hoped that a new test of 140 items could be created. And from a further trial of that test, that a 100 item test could be created for the final version.

Other expectations or hopes arising from a design of this kind are that even using a straightforward traditional analysis, two parallel versions of the final test could be created, thus increasing test security. Strictly speaking items would not be interchangeable individually, since on a traditional analysis any item statistics are 'locked' into the test as a whole. However, for reasons discussed earlier to do with the homogeneity of this particular population, this may be a possibility in this case.

Two forms of the test were thus written – Form A and Form B. These were designed to be parallel tests, but shared no items in common. Clearly, if we are to compare item analyses we need to have more than one set of responses to the items. One way around this problem would be to use high and low scoring sections in the group which takes a particular test and compare their performances on items. This has been tried before (by Woods and Baker (1987) for example); the disadvantage of this method is that it attempts to make traditional statistics do something they were never designed to do, namely to provide stable item statistics. IRT methods, on the other hand, normalise distributions and allow for extremes. It is quite clear that facility values calculated for the high scoring group are by definition going to be much higher than facility values for the same items calculated by using the low scoring group. IRT methods appear to show a more stable item statistic (Woods and Baker op. cit.), but even so we should be cautious: there is evidence that a similar effect, though obscured, is at work with IRT analyses too (Loyd and Hoover 1980). The argument is that by looking at the extremes we can test the models to destruction. This, however, ignores the very different premises upon which IRT and traditional analysis are based. We prefer here to use item statistics derived from whole groups to see if the differences, if any, are really as important as are suggested. Accordingly, two further versions of the test were constructed, Forms C and D, each of which has exactly the same structure and content as Form A and Form B, but which are made up of half the items in A and B combined.

Form C is thus composed schematically of the first half of Form A and the second half of Form B, while Form D is composed of the second half of Form A and the first half of Form B. This, at least, is the conceptually skeleton. In practice, the items are

arranged in such a way as to avoid consecutive runs of content-related items, so the composition of Forms C and D is slightly more complex than this outline would suggest. Forms C and D are given in Appendix II; the exact relationship between Forms A and B and Forms C and D is shown in Appendix III.

A further modification to this design is that in Part 3 of the test (to be discussed shortly) an extra 32 items were used in place of items from Form A and Form B. This was done in order to provide extra material for piloting. No dual item statistics are available, therefore, for a number of Part 3 items. The extra items are shown in Appendix II.

The first version of the test was administered in December 1986 to 1056 first year undergraduate students at USM. Approximately equal numbers took each of the four Forms of the test, which were randomly assigned by the staff at USM on the basis of identity numbers. The students were taken from various departments of the university, and a large percentage had already undergone one semester (45 - 50 hours) of instruction in English at the time they took the test. It is clear that this will distort the test results slightly in so far as some of the test content may have been taught. This is an unavoidable restriction, and any distortion it produces should be ironed out at the second administration of the test.

After analysis of the results of this test (see Chapter 7) a second pair of pilot tests was constructed from the initial 312 items. The two tests had 100 items each. This version was administered in July 1987 to 1,392 students new to USM. Subsequent analysis of this test led to the construction of two 80 item tests, which became the final version of the placement test. The composition of the second pilot tests, and their relation to the original pilot tests are shown in Appendix IV.

6.5. Content validity

In keeping with the principles outlined in Chapter 3 it has already been noted that the intention in writing the test was to ensure a good measure of validity from the start so that any theoretical issues raised could be adequately assessed (particularly in relation to the structure of the dimensions of EFL reading). These principles, if adequate, should also go a long way towards solving the basic placement problem, in that failure on the part of the testee to demonstrate sufficient competence at any one point of the test in relation to content at a particular level will lead to his being automatically assigned to that level. We procede on the basis that the parts of the test should reflect the levels they are designed to relate to. While this principle was discussed in Chapter 3, the principles of item construction itself are based on ideas

outlined in Chapter 4. In order to appreciate the importance of the relationship between test content and course content, we now show briefly the nature of English courses at USM.

6.5.1. Course content: general

Generally speaking, English courses at USM are designed from a 'communicative' methodology standpoint. They have clearly been strongly influenced by *Reading and Thinking in English* (British Council 1980), such that this series of coursebooks forms the basis of much of the work in all English courses. There are also large amounts of extra materials, in the form of course books and specially written supplementary materials, but the spine of the coursework up to Level 3 derives from *Reading and Thinking in English.*

What follows is a summary of the courses available and their overall aims. The information is taken from individual course prospectuses.

6.5.1.1. General English

1. Courses are designed to meet the specialised needs of arts students whose main concern is to understand texts and journals related to their fields of study.

2. Courses aim to develop reading comprehension skills in the context of non-science prose e.g. social and economic issues.

3. Courses aim to develop the communication aspects of reading i.e. to show how language functions as a medium for understanding the writer's views etc.

6.5.1.2. Scientific English

1. Courses are designed to meet the needs of science students to understand modern scientific writing.

2. Courses aim to expose to the students the features of English (lexis, syntax and semantics) occurring frequently in scientific writing.

3. Courses aim to develop reading comprehension skills in the context of science.

4. Reading as a communicative ability is developed.

6.5.1.3. English for Mass Communications

1. To enable students to develop and acquire the basic skills needed to read and study effectively in English at undergraduate level.

2. To enable students to develop the aspects of grammar which facilitate speaking, reading, and writing tasks.

3. To enable students to develop reading-integrated writing skills.

6.5.1.4. Business English

1. To develop basic reading comprehension skills

2. To enable students to approach reading as a communicative process.

3. To expose students to authentic texts from articles and management journals.

4. To enable students to develop the aspects of grammar when necessary which facilitate speaking, reading and writing tasks.

6.5.1.5. Technical English

1. To assist students to become proficient readers of academic and occupation-related texts in the technical and engineering disciplines.

2. To develop oral and written skills.

6.5.1.6. English for Building and Planning

1. To enable students to acquire and develop the basic reading skills for effective comprehension of a range of language functions.

2. To enable students to acquire a basic understanding of grammatical structures to facilitate their reading etc.

6.5.2. Course content at different levels

Each Level consists of 56 hours of teaching i.e. one semester with two two-hour classes per week.

6.5.2.1. Level 1

1. General English: working knowledge of English grammar and vocabulary; parts of speech etc.; short reading selections.

2. Scientific English: familiarisation with common rhetorical functions; words in context; cause and effect, purpose and method etc.; main idea and supporting details. Set book: *Reading and Thinking in English Book 2*

3. Mass Communications: rhetorical functions; skimming, scanning etc.

4. Business English: functions of texts; specific reading skills.

5. Technical English: functions of texts; grammar.

6. Building and planning: time, location; rhetorical functions; basic reading skills such as transfer of information from linear to non-linear form; words in context.

6.5.2.2. Level 2

1. General English: logical relationships; text functions through e.g. transitional markers and connectives; specific reading skills such as identifying main ideas, inferences, conclusions; reading as a thinking process; vocabulary building.

2. Scientific English: reading comprehension; grammatical areas common in scientific texts e.g. relative clauses, passive verbs, connectors; text function; words in context.

3. Mass Communications: use of linking words/phrases, cause-effect relationships, comparison/contrast, non-linear text; vocabulary in context; referring phrases; text functions e.g. classification.

4. Business English: text organisation; extracting information from text; comparisons, causal relationships, purpose and other rhetorical functions; linking words; information transfer (linear to non-linear text and vice- versa).

5. Technical English: assessing relevance at whole-book and Chapter level; rapid reading for gist of short texts; using textual clues; functions of technical English such as cause-effect relationships.

6. English for Building and Planning: referencing skills at whole-book and Chapter level; rhetorical funtions e.g. process descriptions; greater lexical and structural complexity.

6.5.2.3. Level 3

1. General English: wide variety of relevant texts - fiction and non-fiction; rhetorical functions e.g. main ideas, inferences, cause and effect, forming conclusions.

2. Scientific English: longer, more complex, more abstract texts; emphasis on overall comprehension of texts e.g. distinguishing facts from opinions; logical relationships between words, sentences and paragraphs; understanding of scientific ideas using explanatory techniques; scientific terminology through word components; mechanics of reading including eye-fixation.

3. Mass Communications: stylistic and rhetorical elements in texts e.g. topic sentences, transitions, reference; skimming, scanning; authentic reading (journals); reading as a thinking process; assessing main ideas.

4. Business English: higher order reading comprehension skills; skimming, scanning; predicting, inferencing, hypothesising, distinguishing between fact and opinion; authentic reading (magazines).

5. Technical English: overall organization of texts; inter-sentential functional relationships in a text; vocabulary in context; cohesive links in a text; connectors.

6. Building and Planning: texts complex in terms of lexical-structural load; critical evaluation of texts; unravelling of the organisation of texts in terms of functional value e.g. comparison, cause-effect etc.; distinguishing fact from opinion; assessing degrees of certainty.

NB There is a seventh course for 'Matriculation' students i.e. students who have finished school and are waiting to enter the university; they are nearly all science students and while the course is described in similar terms to those used above, it has a much more overt reliance on set texts. The set text for Level 1 is *Reading and Thinking in English Book 1*, for Level 2 is *Reading and Thinking in English Book 2* and for Level 3 *Reading and Thinking in English Book 3*

6.5.2.4. Relative numbers

The proportion of students taking any of these courses is relatively stable from year to year. The figures for 1986 are typical:

Table 1
Numbers of students on USM courses /cont.

Table 1
Numbers of students on USM courses

	Total	Level 1	Level 2	Level 3	Level 4
General	1000	213	198	349	240
Science	624	20	97	213	294
Medicine	|	0	4	36	32
M. Comm.	|	32	19	39	0
Business	501	26	31	37	10
Technical	|	3	36	69	40
Building	|	13	26	27	21

6.5.3. Comments on course content

What may not be clear from the above course descriptions is the extent to which in practice most of the courses follow the basic outline of *Reading and Thinking in English*, as the matariculation course makes plain. Nor will it be clear that in fact many of the differences between these courses relate simply to supplementary reading material. This is not a criticism of the courses, since many of them are new and will doubtless develop their own characters in time, but it should be emphasised that for placement testing in the situation as it is at present and is likely to continue to be for some time yet, there is not such a large difference in content as might at first appear. This is a further argument, of course, for restricting the test versions to one for all students.

Another feature which is not immediately apparent from the above description is that in fact General English is about one level below the other levels. By this we mean that Level 2 General English corresponds roughly to Level 1 Scientific English and so on. Level 1 General English is a much more basic course than the description suggests. The reasons for this are not immediately clear - perhaps science students have had to read more texts in English throughout their education, whereas Arts/General students have, perhaps ideologically, turned away from English texts as a matter of principle. Whatever the reason, the implication is that althought the same test can be administered to all students, placement criteria may have to be different. This would not be a serious problem were we not trying to construct a three-part test which reflects the course content of the three levels. Our decision in this case is based upon the reasoning outlined earlier that the costs of misplacement at the lower levels are perhaps less significant than the costs of misplacement at the higher levels. Though it is always invidious to have to make such decisions, the exigencies of the situation are such that we must decide our emphasis at this point. Accordingly it was

decided to take the Matriculation/Scientific English courses as the 'prototype'. Once an item bank has been created and is working this should prove less of a problem, since a test of suitable length could be tailored for arts and science students. At this stage, however, the sheer length of a test which would have to discriminate at, effectively, four levels was felt to be prohibitive. This will be a confounding factor, however, when we come to discuss the results later.

The syllabus and the content descriptions of the courses provide a kind of list of specifications (in the Munby 1978 sense) and could therefore be used as a basis for item-writing. The problem with this approach is that while it is perfectly possible to use a list of specifications as an aid to item construction - an *aide-memoire* perhaps - the process falls down when applied in reverse, and that is, after all, the real point of interest.

To look at this point in more detail, consider the following argument. Given a list of specifications it is relatively easy to construct an item. One may then say that the item tests, for example, "understanding words in context". THe problem arises that, unless there is some glaring mismatch such as between the specification "skimming" and the item "Fill in the blank: This camera _____ good pictures", then one can *say* that the item tests whatever one wants it to. This was part of the problem with the analysis of subskills discussed in Chapter 2. Take again the specification "understanding words in context". An item such as "*He was so parthenophobic that he lived in the red-light district.* What does *parthenophobic* mean?" could be said to be testing "words in context" for some readers. But it would be equally possible to say, given only the item, that it is a test of "inferencing" (another favourite skill, separable from "understanding words in context") or even a test of "cultural knowledge". It could certainly, for some readers, be a test of vocabulary/"word knowledge" (a separate category in most taxonomies, though not in Munby, from "understanding words in context"). And this is a fairly simple example; the point is that the validity of the item (both in terms of content and construct) is not guaranteed, or even strongly aided, by working from a list of specifications alone, however exhaustive. In slightly different terms one could say that the route to an answer, the processing path which the testee takes, could be different from one individual to the next. This is the argument against relying too heavily on needs specifications (cf. Criper and Davies 1986; section 8.2).

What is needed is the addition of a domain from which test items can be created (as argued in Chapters 3 and 4). This is not always available or easy to obtain; for example, in a general language proficiency test, while the domain is certainly there, it

is very difficult to limit its boundaries with sufficient precision to be useful for the test-writer's purpose. The proof that a domain, however unclear, exists can be readily appreciated when one considers the kind of texts which would be considered suitable as reading comprehension passages for, say, Cambridge Proficiency: why would a 2000 word extract on yak farming be unsuitable? (Though it is worth noting that extracts from knitting patterns were occasionally included in the older tests.) However, in this case the domain is defined with almost exhaustive clarity. The text-book for a particular course, as was noted in Chapter 4, could be considered a domain. The problem really is one of selection, which is the question we now tackle.

6.5.4. Test content: Part 1

The items for this part of the test are designed to reflect the course content at Level 1. This is a very basic level and concentrates almost entirely on what might be called 'grammar'. This is an unfortunate term since it tends to arouse adverse reactions in more 'enlightened' teachers and testers; for USM however, as was suggested above, "basic grammar" is an essential component of reading. Our discussion in Chapter 2 further suggested that for foreign language readers at any rate (if not for second language readers) the 'grammar' element of reading will be ignored at the learner's peril. A case could also be made for considering the 'grammar' component of a test as a test of a 'low-level' reading skill. For reasons of face validity, Part 1 of the test was called a test of Basic Grammar; for a different audience it might be re-labelled Lower Order Reading or something similar. We return to the problem of labelling items in Chapter 8.

Using domain-sampling techniques outlined in Chapter 4 it was found that Level 1 work concentrates basically on the sentence, more or less in isolation, and covers ten major content areas defined in traditional grammatical terms. The areas, together with sample test items, are as follows:

1. Past/Perfect Verb Forms: *I hope you are enjoying your visit; _____ any new friends yet?* a. did you make b. are you making c. have you made d. do you make (item B37)

2. Present/Future Verb Forms: *Don't leave the aeroplane until the steward _____ you to.* a. will tell b. would tell c. is telling d. tells (C5)

3. Conditionals: *If I had a lot of money I suppose _____ buy a house.* a.I'd buy b.I have bought c.I'll buy d.I'm buying (D3)

4. Prepositions: He tried to drive away the wolves by throwing stones ____ them.) a.to b.in c.for d.at (B1)

5. Conjunctions: *We can still go and see him ____ it is raining*) a.even though b.whereas c.besides d.despite (C11)

6. Relative Pronouns: *The car crashed into a queue of people _____ four were killed.* a.where b.so c.of whom d.by which (D6)

7. Articles: *A concerto is _____ piece of music for orchestra and solo instrument.* a.the b.some c.a d.____ (B5)

8. Modals/Auxiliaries: *This play _____ written by Shakespeare because the style is Milton's.* a.was b.couldn't have been c.could have been d.need have been (C4)

9. Infinitives/Gerunds: *It is better _____ your money in a bank than under the bed.* a.to put b.putting c.by putting d.put (D1)

10. Comparisons: _____ *animals, plants do not move.* a.like b.as c. by comparison d.unlike (B31)

5 items were constructed for each category, thus giving a total of 50 items for Part 1. The exact item numbers corresponding to these 10 categories are given in Appendix I.

A total of 100 items was thus available for piloting for Part 1, with two independently derived sets of item statistics for each item.

6.5.5. Test content: Part 2

Items for this part are designed to reflect the course content at Level 2. Essentially what we find is a kind of sentence-level, paraphrase type of task, similar to the first section of the ELTS General Reading Test (G2). A fair amount of metalinguistic knowledge is also introduced at this level, and while it may be argued that testing metalinguistic knowledge is unfair in a test of general proficiency for incoming candidates in so far as they may be able to read perfectly well and yet still not understand the metalinguistic terminology employed, it can also be argued that if a test is required to place students on a programme where such knowledge is to be taught then the students will be disadvantaged if they are not familiar with the terms.

Thirteen broad areas of content can be distinguished at Level 2, as follows:

1. Active/Passive: *This speed limit is to be introduced gradually.* a.They will introduce this speed limit gradually b.This speed limit which will be introduced is very low c.The gradual introduction of the speed limit will be done soon d.The gradual introduction will lower the speed limit (B57)

2. Cause/Effect: *He works too fast; that's why he makes so many mistakes.* a.If he didn't work so fast he wouldn't make so many mistakes b.If he worked faster he wouldn't make so many mistakes

c.If he works too fast he'll make so many mistakes d.If he had worked faster he would have made so many mistakes (B70)

3. Purpose/Result: *Shelters have been built in case of war breaking out.* a.Due to the outbreak of war, shelters have been built b.Shelters have been built in order to prevent war breaking out c.The outbreak of war has led to the building of shelters d. If war breaks out shelters will already have been built (B54)

4. Reported Speech: *I said: "Let's not jump to conclusions. Let's wait till we hear confirmation of this rumour."* a.I told everyone not to jump to conclusions b.I ordered everyone not to jump to conclusions c.My advice was to wait till we heard confirmation of the rumour d.My suggestion was to wait unless we heard confirmation of the rumour (B52)

5. Similarity/Difference: *I dislike flying in the way that you dislike sailing.* a.I don't like flying and neither do you b.I don't like sailing or flying; nor do you c.I like neither flying nor sailing, like you d.I don't like flying, and you don't like sailing (B65)

6. Sequence of events: *Prior to his return he had meant to throw it away.* a.His intention was to throw it away before he returned b.He intended to return and then throw it away c.On his return he intended to throw it away d.When he had returned his intention was to throw it away (B51)

7. Relative Clauses: *He didn't thank us, which offended us.* a.He must have thanked us b.He needn't have thanked us c.He should have thanked us d.He didn't need to thank us (B62)

8. Connectors: *Walk carefully over the floor. Otherwise you may fall.* a.If you don't walk carefully you won't fall b.Unless you walk carefully don't fall c.Walk very carefully unless you fall d.Walk very carefully lest you fall (B55)

9. Possibility/Certainty: *It looks as if food will soon be cultivated under the sea.* a.We think that soon food may possibly cultivated under the sea b.We think that food is unlikely to be cultivated under the sea c.The cultivation of food under the sea is an unlikely prospect d.The cultivation of food under the sea is now a realistic possibility (B56)

10. Text types: *A device used for keeping food fresh or frozen is called a refrigerator.* Is this sentence a.classifying b.defining c.describing d.exemplifying? (B)

11. Logical Ordering: *a.Despite its wide distribution, carbon constitutes only 0.19% of the earth's crust. b.Carbon is a solid non-metallic chemical element occurring in the pure chrystalline form as diamond and graphite. c.It is also found in the combined form as a constituent of all organic materials, including coal and petroleum.* Which is the first/second/third sentence? (B78)

12. Transfer of information: *see Appendix II items 85 - 94 on either Form.*

13. Connectors in Discourse: *Cultures have definite patterns. (B95* these patterns are modified (B96) they are transmitted from one generation to the next.) (B95) a. Although b.But c.Since d.Also (B96) a. as b.during c.also d.so

Each Form of the test has 50 items in Part 2, the distribution of which is shown in Appendix I.

6.5.6. Comments on test content: Parts 1 and 2

The further one moves away from tightly controlled items the less easy it becomes to say, for example, that such-and-such an item is a test of 'relative clauses'. The example from Part 2 given above, for example, certainly tests an understanding of modals as well - perhaps even more than it tests an understanding of relative clauses. This means that a rigid separation of content into different areas of this kind for this kind of test will just not be possible.

One of the implications of this is that a fully-fledged diagnostic model may be inherently flawed. This will be further discussed in Chapter 7.

6.5.7. Test content: Part 3

This part of the test is meant to reflect work at Level 3. This means the exercise of reading skills on extended text (often very extended!) It was decided to make Part 3 entirely in the format of a reading comprehension test. The choices we have to make at this stage concern (i) the number of texts to include (ii)the length of the texts (iii)the number of items to be included with each text (iv)item types. Note that in using our current domain-sampling approach we tend to by-pass the issue of textual difficulty; all we commit ourselves to in the current format is to saying that somehow extended texts are more worthy tests of advanced reading than single sentences. This fits in with our two-tier view and means that we do not have to enter the difficult realm of comparing difficulties of texts (though for interest and for rough comparisons readability indices will be reported later).We have argued all along that to talk of difficulty of texts is slightly misleading in that it is fairer to say, for testing purposes certainly, that difficulty is a function of the task (cf. discussion in Chapter 4). However, there are other difficulties with this kind of test: in the interests of greater reliability it was decided to use a variety of texts rather than one or two long ones. To comply with the requirements of typical texts at this level it was decided 5 texts should be sufficient. As far as the number of items is concerned, we argue that too many items on one text will have the effect of splitting the text too much and rendering the type of reading involved more akin to what was hoped would be tested

in Part 2 i.e. relationships between one or two sentences. 8 items per text was felt to be sufficient.

As far as item types are concerned, we discussed in Chapter 2 the merits of thinking in terms of subskills, and tended to reject that idea. However, in terms of test construction, and because the teaching is so structured, we should attempt to match items to 'skills' as they are perceived. This, it should be noted, is different from the approach in Parts 1 and 2: there it was possible to have a list of specifications *and* a domain of content from which to sample. At Level 3 the domain tends to be more difficult to arrive at. Certainly there is no problem so far as the actual texts are concerned - there is a readily identifiable domain of texts. However, the items are not inherent parts of the text in the way that items in the present tests' Parts 1 and 2 are in most cases very closely related to the domain. True domain sampling would involve using text and task together (*pace* Hively 1974), but most tasks at Level 3 in instructional use are not really appropriate for or easily adapted to the computer markable format. A case in point would be if one wanted to test "scanning": the whole point about scanning is that it is done quickly for a particular purpose; typically there would be one piece of information for which one would be searching a particular text. Now in a classroom this kind of activity is possible because one can either use the text *only* for the skimming exercise and then move on to another text or use the text for the skimming exercise and *then* use the text for some other purpose - the skimming, one hopes, will have been effectively carried out. But in a testing situation one does not have either of these possibilities: it is wasteful of time and resources to use a large text for one question - this is simply inefficient testing which will also in all likelihood reduce reliability. The usual way of overcoming this problem is to put the 'skimming' question at the head of the queue as it were and then to go on and ask other questions. But this does not take into account the fact that once a text has been looked at in some detail (as presumably it must be if there are several questions on it) then it is no longer possible to skim-read it in the manner intended. Test-wise candidates will be able to go back to the 'skimming' question and change their answer (if they wish). A similar argument holds for 'predictive' question types: they are easily used in teaching but very difficult to use in testing.

We should, then, feel less sure of our ground in matching items to specifications in the case of reading comprehension. The method we adopt here is to select a large number of texts currently in use at Level 3 (in fact 24 texts were originally selected), write items for them using principles to be outlined shortly, and then select the 14 which would be used in the pilot version. In a few cases texts and items were already in the course and could be used with very little adaptation. Most however

had to have items written for them.

The specifications used to write these items were those common skills identified by Rosenshine (1980) and discussed in Chapter 2. This does not mean that we commit ourselves to strong support of the 'subskills' hypothesis, nor indeed that there are not equally compelling reasons for choosing other lists of item types, merely that a good test of reading comprehension should contain a variety of item types, the scores on which, when pooled, will provide an indirect measure of a testee's reading comprehension ability (cf. Lunzer *et al.* 1979)

The exact correspondence of items to specifications is given in Appendix I.

6.5.8. Comments on test content: Part 3

The information contained in Appendix I can be reduced to show the balance of 'subskills' being tested:

Table 2
Subskills tested in Part 3 – all Forms

Subskill	No. of items
Recognising sequence	2
Recognising words in context	19
Identifying the main idea	22
Decoding detail	14
Drawing inferences	30
Recognising cause and effect	16
Comparing and contrasting	9

Clearly there is an imbalance in the skills tested; however, since the questions that can be asked of a text depend to a large extent on the text itself (some texts simply do not support certain types of question) this would seem to be an unavoidable problem.

6.6. Method of analysis

6.6.1. Traditional statistics

All test results were first analysed using a traditional item analysis procedure (using a Pascal computer program – Tradanal – written by Alistair Pollitt of Edinburgh University's Godfrey Thompson Unit). For the immediate purposes of the placement test project, the 200 best performing items were selected to produce two further versions for subsequent trials.

6.6.2. IRT analysis

A Rasch analysis of the test was carried out using the computer program BICAL (Wright and Bell 1981). Each form was analysed first as a whole (i.e. on the assumption that one scale could account for all the data) and then with respect to each of its sections (on the assumption that each Part represented a separate dimension and would thus require a separate scale). Analysis of fit was also carried out.

6.6.3. Validity and dimensionality

The construct validity of the test was investigated using Factor Analysis (SPSS–X) and the concurrent validity was examined by comparing test results with other available measures of student ability, namely test results in either of the two school leaving certificates (SPM 322 and SPM 1119) and current English level with respect to course level at USM. The factor analysis was used as the basis for an investigation of the dimensionality of the set of reading items.

6.6.4. Comparison of Traditional and Rasch Analyses

The two methods of analysis were compared to see what advantages one had over the other for a situation of this kind.

6.6.5. Difficulty and ability

Different ways of arriving at estimates of difficulty and ability were investigated, and an attempt was made to relate the statistical findings to the analysis of content in order to arrive at some understanding of what we might want to understand by 'difficulty' and 'ability' in this context.

6.6.6. Item bank construction

Each of the items in Forms C and D was calibrated from the linking items first in Form A and then in Form B. The results were compared with the actual values obtained on the administration of Forms C and D. The design of the test enabled a comparison to be made between the use of different sets of anchor items, thus establishing whether an item bank could be confidently extended using data of this kind.

CHAPTER 7

ANALYSIS AND DISCUSSION OF RESULTS

7.1. Introduction

The analysis and discussion which is presented in this chapter is designed to answer the following questions:

1. How do the test data perform under traditional analysis?
2. How do the Rasch statistics fit the data?
3. How do the traditional and Rasch analyses compare?
4. How many dimensions are there to the test data?
5. What do the test statistics reveal about the difficulty of the test items from a content point of view?
6. How can 'ability' and 'difficulty' be related to form a preliminary item bank?
7. How do calibrations of all items on a single scale compare if different anchor sets are used?
8. Is item banking a workable concept with data of this kind?

7.2. Analysis of test results: classical and Rasch statistics

7.2.1. Summary of classical descriptive statistics

The full item analysis of the four forms of the test can be found in Appendix III (full details of the item-by-item responses are held by the present author and by the Language Centre at USM, from either of whom further information may be obtained). The following summary shows the main implications of the test results from the traditional point of view. The reader should bear in mind that Part 1 is a single sentence completion grammar-type test, Part 2 is a paraphrase recognition/short discourse interpretation test, Part 3 is a comprehension-type test, and Part 4 is a summary writing test. (see Appendix II) The Part 4 results are included where this is possible and relevant to the discussion.

The maximum possible score obtainable is 50 for each of Parts 1 and 2, 40 for Part 3, and 30 for Part 4 - a total of 140 if Part 4 is excluded, or of 170 if it is included. Inspection of the table below will reveal that in general students taking Form A consistently found all parts of the test easier than those taking any of the other forms

– they demonstrate a higher 'ability' level than other groups. The other three groups do not appear to be very different from each other in terms of ability.

Part 2 of Forms A and C have higher means than for their Part 1. This is an unexpected and unintended result, as one assumption in the design of the test was that Part 2 would consistently be more 'difficult' than Part 1, if for no other reason than that it contains more reading matter. A closer look at the item analysis shows that a number of the items in Part 2, particularly those relating to Information Transfer and to Logical Ordering have very high facility values; this goes some way towards explaining the imbalance in mean scores. Another, related, factor may be that the interdependence of several of the Part 2 items, particularly those that appear to be 'easier' than the rest, is contributing to a distortion which is caused by the test analysis treating as separate ('locally independent') items which in fact should be treated together.

This is an important observation, since it shows that while classical test statistics are in general more robust than IRT statistics in that they do not make strong assumptions about the data (e.g. that items are locally independent), nevertheless the existence of items which perform 'together', as it were, can affect the interpretation of classical analysis. One implication of this is that several of the assumptions underlying IRT are in fact already assumed, tacitly, in a classical analysis.

For all four forms of the test, Part 3 (reading comprehension) has a lower mean (proportionately) than Parts 1 and 2, while Part 4 has a lower proportionate mean than all the other parts. This follows the expectation built into the test that somehow summary writing is more demanding or 'difficult' than reading comprehension, which in turn is more 'difficult' than paraphrase recognition or basic reading grammar. At the same time it should be noted that the Part 3 means in Forms B and D are markedly lower than those in Forms B and D, which suggests that the reading comprehension items in Forms A and C were 'easier' for the students taking those tests.

Further evidence for the relative easiness of Forms A and C can be found in the higher means for the total score in those forms. It could be said, however, that the differences are too small to take into account anything other than the fact that the Part 3 scores differ i.e. that the difference in Total score is a duplication of the difference in Part 3 score.

So far as range of ability is concerned, an inspection of the standard deviations of the four Forms shows that Form C disperses students the most, thus we could

conclude that the group taking this Form of the test exhibited the greatest variation in ability. The variation among students as shown by the standard deviations of the sub-parts of the various forms does not exhibit any unusual pattern. The only point of interest is that in Parts 3 and 4 of all Forms, the standard deviation is large in relation to the mean, suggesting that these Parts possess greater discriminating power than the other Parts.

The high reliability coefficients for Parts 1 - 3 of all Forms is very satisfactory, especially for the Part 3 sections. The slightly lower reliabilities reported for Part 2 in all Forms when compared with Part 1 provides slight evidence for the existence of heterogeneity in test content in that Part, in so far as a measure of internal consistency such as is reported here relies on homogeneity of content. This point will be explored further when we look at the construct validity of the test.

All items which discriminated at the 0.3 level and above were retained for inclusion in the second pilot version of the test. The results of the second pilot are not reported here in detail; readers wishing to examine the results of that administration of the test should seek further information from the present author, from USM or from The British Council.

The most important feature of the item analysis statistics lies in the results for Part 2. The items on text types were found to discriminate well in one Form of the test but not to discriminate at all (or to discriminate negatively) in the other Form of the test in which they appear in exactly the same format. This odd result will be considered in more detail in the discussion of construct validity.

The full summary of results is as follows:

Table 3
Summary statistics for Forms A - D

FORM A

	N	MEAN	S.D.	KR-20
PART 1:	269	32.7	7.8	0.87
PART 2:	269	36.7	6.5	0.85
PART 3:	269	19.2	6.1	0.79
PART 4:	237	11.3	4.3	
TOTAL:	269	98.6	21.1	0.93

FORM B

	N	MEAN	S.D.	KR-20
PART 1:	245	32.7	10.0	0.92
PART 2:	245	29.4	7.9	0.85
PART 3:	245	16.9	5.9	0.77
PART 4:	239	10.9	4.1	
TOTAL:	245	89.7	23.7	0.95

FORM C

	N	MEAN	S.D.	KR-20
PART 1:	276	29.4	9.3	0.90
PART 2:	276	33.2	7.7	0.87
PART 3:	276	19.3	6.7	0.83
PART 4:	253	10.6	4.1	
TOTAL:	276	91.7	24.4	0.95

FORM D

	N	MEAN	S.D.	KR-20
PART 1:	266	31.3	8.9	0.89
PART 2:	266	28.1	7.5	0.82
PART 3:	266	16.6	6.1	0.79
PART 4:	241	11.3	5.2	
TOTAL:	266	86.3	23.5	0.94

7.2.2. Stability of classical item statistics

The correlation between facility values obtained in Form A of the test with those obtained for the same items in either Form C or D is 0.94. The correlation between facility values obtained in Form B of the test and those obtained for the same items in either Form C or D is also 0.94. This appears to show that the facility value is not such a volatile index as is sometimes claimed. On the other hand it must be remembered that the population here is a relatively homogeneous one and so one would not expect startling variations in the classical indices.

The correlation between facility values obtained in the first pilot test and those obtained in the second pilot test is 0.93 (for second pilot Form A) and 0.82 (for second pilot Form B). This confirms the tentative conclusion of the previous paragraph, and indeed is perhaps more significant in view of the fact that the second test-taking population did not share the USM educational background that the first population had.

7.2.3. Rasch analysis: statistics for the whole test

The first column after the item label in the following tables shows the Rasch estimate of item difficulty for each item as if that item were part of a single 140-item test - in other words, no account is taken of whatever dimensions might be present (see section 7.3 for a fuller analysis of the difficulty estimates in terms of dimensions). To a certain extent this is to pre-empt conclusions which will be made later. The difficulty estimates reported here are UCON estimates. PROX estimates were also obtained, but these are slightly less accurate in that they use less information to arrive at a result. Nevertheless, the correlation between UCON and PROX estimates was above 0.92 for all forms of the test.

The tables are organised so that the first panel shows the item statistics in item serial order, the second panel shows the item statistics in order of increasing item difficulty (e.g. in Form A the 'easiest' item is item 94, the most difficult is item 76), while the third panel shows the item statistics in order of decreasing goodness of fit (thus in Form A, the best fitting item is item 17, while the worst fitting item is item 110). These results will be discussed in the next section.

Table 4
Full statistics from Rasch analysis

SEQ NUM	ITEM NAME	ITEM DIFF	STD ERROR	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	INFO AMT	FIT T-TESTS BETWN	FIT T-TESTS TOTAL	WTD MNSQ	MNSQ SD	DISC INDX	POINT BISER
1	A 1	0.24	0.15	1.13	-0.84	94	A 94	-4.00	1.93	0.11	17	A 17	1.01	47	2.44	-3.85	0.84	0.04	1.69	0.54
2	A 2	-1.63	0.26	1.79	-0.44	21	A 21	-3.59	0.79	0.13	13	A 13	0.45	14	2.80	-3.75	0.82	0.05	1.83	0.57
3	A 3	-1.53	0.23	1.37	-0.22	53	A 53	-3.07	1.34	0.06	114	A114	1.51	18	2.95	-3.55	0.83	0.05	1.68	0.53
4	A 4	-0.62	0.17	0.20	0.98	89	A 89	-2.88	0.30	0.01	121	A121	1.74	33	2.30	-3.10	0.83	0.06	1.75	0.53
5	A 5	1.44	0.14	0.67	1.61	93	A 93	-2.88	1.54	-0.05	46	A 46	1.36	50	1.82	-2.84	0.87	0.05	1.59	0.49
6	A 6	-0.85	0.19	1.22	-0.05	58	A 58	-2.88	1.52	0.03	70	A 70	0.95	29	0.90	-2.53	0.89	0.04	1.48	0.48
7	A 7	-2.34	0.32	1.14	-0.03	91	A 91	-2.72	0.89	0.11	116	A116	1.66	9	1.39	-2.45	0.87	0.05	1.56	0.49
8	A 8	0.39	0.14	0.78	0.97	56	A 56	-2.72	1.17	0.06	27	A 27	0.53	48	1.00	-2.30	0.89	0.05	1.47	0.47
9	A 9	-1.64	0.24	1.67	-0.57	90	A 90	-2.72	0.73	-0.02	73	A 73	0.18	16	1.37	-2.22	0.87	0.06	1.59	0.49
10	A 10	-1.70	0.25	1.47	-0.21	98	A 98	-2.72	1.76	-0.10	40	A 40	0.24	16	2.05	-2.19	0.88	0.06	1.53	0.48
11	A 11	-1.98	0.28	1.78	-0.29	96	A 96	-2.58	1.59	-0.16	63	A 63	1.18	12	1.31	-2.14	0.91	0.04	1.38	0.45
12	A 12	-0.42	0.17	1.08	0.19	7	A 7	-2.34	1.14	-0.03	61	A 61	-0.17	36	3.66	-2.00	0.86	0.07	1.79	0.50
13	A 13	0.45	0.14	1.83	-3.75	62	A 62	-2.34	0.97	-0.05	119	A119	-0.14	49	2.28	-1.88	0.87	0.07	1.73	0.49
14	A 14	1.44	0.14	0.58	2.15	24	A 24	-2.34	1.77	-0.07	106	A106	0.04	50	1.05	-1.72	0.89	0.06	1.52	0.46
15	A 15	3.06	0.20	1.07	-0.08	32	A 32	-2.24	1.20	-0.01	136	A136	0.72	25	0.98	-1.68	0.93	0.05	1.29	0.43
16	A 16	-1.58	0.24	1.23	-0.32	92	A 92	-2.24	0.82	0.00	54	A 54	-0.72	17	1.88	-1.55	0.84	0.11	1.72	0.49
17	A 17	1.01	0.14	1.69	-3.85	88	A 88	-2.06	0.63	0.09	60	A 60	0.62	52	0.74	-1.54	0.93	0.05	1.35	0.43
18	A 18	0.22	0.15	1.02	-0.10	49	A 49	-1.98	1.22	-0.07	57	A 57	0.95	46	3.00	-1.38	0.94	0.04	1.26	0.40
19	A 19	-1.32	0.22	1.52	-0.60	11	A 11	-1.98	1.78	-0.29	79	A 79	-0.29	21	-0.34	-1.08	0.91	0.08	1.30	0.41
20	A 20	-1.32	0.22	1.28	-0.08	72	A 72	-1.90	1.10	-0.06	42	A 42	-0.27	21	0.77	-1.05	0.92	0.08	1.45	0.42
21	A 21	-3.59	0.58	0.79	0.13	52	A 52	-1.83	1.61	-0.38	139	A139	1.62	3	1.44	-1.05	0.94	0.05	1.26	0.37
22	A 22	0.41	0.14	1.04	-0.33	44	A 44	-1.83	1.50	-0.21	80	A 80	-0.27	48	-0.02	-1.03	0.92	0.08	1.29	0.40
23	A 23	-0.40	0.16	1.09	-0.13	2	A 2	-1.83	1.79	-0.44	26	A 26	0.68	37	-0.34	-0.96	0.96	0.05	1.22	0.40
24	A 24	-2.34	0.32	1.77	-0.07	45	A 45	-1.77	0.86	0.10	138	A138	0.64	9	0.11	-0.96	0.96	0.05	1.22	0.39
25	A 25	1.28	0.14	0.82	0.81	10	A 10	-1.70	1.47	-0.21	124	A124	1.45	51	0.73	-0.92	0.95	0.05	1.21	0.38
26	A 26	0.68	0.14	1.22	-0.96	68	A 68	-1.70	1.49	-0.26	115	A115	0.15	51	0.36	-0.88	0.95	0.06	1.21	0.39
27	A 27	0.53	0.14	1.47	-2.30	29	A 29	-1.70	1.53	-0.38	1	A 1	0.24	50	1.68	-0.84	0.95	0.06	1.13	0.38
28	A 28	-0.81	0.18	1.55	-0.71	9	A 9	-1.64	1.67	-0.57	34	A 34	2.09	29	2.25	-0.84	0.94	0.07	1.30	0.37
29	A 29	-1.70	0.25	1.53	-0.38	35	A 35	-1.58	1.58	-0.37	81	A 81	-0.19	16	-0.25	-0.82	0.94	0.08	1.23	0.39
30	A 30	1.40	0.14	0.50	2.37	16	A 16	-1.58	1.23	-0.32	66	A 66	-1.32	50	1.93	-0.82	0.87	0.16	1.69	0.41
31	A 31	0.11	0.15	0.85	0.76	85	A 85	-1.53	0.60	0.28	39	A 39	1.49	45	2.10	-0.80	0.96	0.05	1.21	0.37
32	A 32	-2.24	0.31	1.20	-0.01	3	A 3	-1.53	1.37	-0.22	33	A 33	-0.95	10	1.21	-0.79	0.90	0.12	1.64	0.40
33	A 33	-0.95	0.19	1.64	-0.79	82	A 82	-1.47	1.54	-0.26	107	A107	1.68	27	1.24	-0.79	0.96	0.06	1.10	0.36
34	A 34	2.09	0.15	1.30	-0.84	64	A 64	-1.37	1.41	-0.16	95	A 95	-0.27	42	0.10	-0.79	0.94	0.08	1.18	0.37
35	A 35	-1.58	0.24	1.58	-0.37	66	A 66	-1.32	1.69	-0.82	36	A 36	-0.56	17	0.71	-0.75	0.93	0.10	1.42	0.37
36	A 36	-0.56	0.17	1.42	-0.75	19	A 19	-1.32	1.52	-0.60	100	A100	-1.07	34	1.26	-0.73	0.90	0.13	1.54	0.38
37	A 37	1.34	0.14	0.54	1.83	86	A 86	-1.32	0.16	0.51	28	A 28	-0.81	51	0.70	-0.71	0.92	0.11	1.55	0.39
38	A 38	1.89	0.15	0.29	1.68	20	A 20	-1.32	1.28	-0.08	108	A108	0.47	45	0.87	-0.69	0.97	0.05	1.11	0.38
39	A 39	1.49	0.14	1.21	-0.80	122	A122	-1.32	1.42	-0.40	127	A127	1.22	49	1.65	-0.67	0.97	0.05	1.14	0.38
40	A 40	0.24	0.15	1.53	-2.19	118	A118	-1.28	1.26	-0.40	47	A 47	1.62	47	-1.12	-0.60	0.97	0.05	1.14	0.36
41	A 41	3.67	0.24	1.20	-0.28	100	A100	-1.07	1.54	-0.73	19	A 19	-1.32	16	0.41	-0.60	0.90	0.16	1.52	0.37
42	A 42	-0.27	0.16	1.45	-1.05	67	A 67	-0.95	0.81	0.41	9	A 9	-1.64	39	1.58	-0.57	0.88	0.19	1.67	0.36
43	A 43	0.32	0.14	0.24	2.83	33	A 33	-0.95	1.64	-0.79	120	A120	-0.62	48	-0.54	-0.53	0.94	0.10	1.25	0.36
44	A 44	-1.83	0.26	1.50	-0.21	78	A 78	-0.88	0.89	-0.09	2	A 2	-1.83	14	1.59	-0.44	0.90	0.21	1.79	0.35
45	A 45	-1.77	0.26	0.86	0.10	6	A 6	-0.85	1.22	-0.05	118	A118	-1.28	15	0.15	-0.40	0.93	0.15	1.26	0.30
46	A 46	1.36	0.14	1.59	-2.84	87	A 87	-0.85	0.14	1.03	122	A122	-1.32	50	0.35	-0.40	0.93	0.16	1.42	0.33
47	A 47	1.62	0.14	1.14	-0.60	28	A 28	-0.81	1.55	-0.71	29	A 29	-1.70	48	0.84	-0.38	0.91	0.20	1.53	0.32
48	A 48	0.76	0.14	1.05	0.17	104	A104	-0.75	0.58	0.36	52	A 52	-1.83	51	1.49	-0.38	0.91	0.21	1.61	0.32
49	A 49	-1.98	0.28	1.22	-0.07	54	A 54	-0.72	1.72	-1.55	35	A 35	-1.58	12	0.47	-0.37	0.92	0.18	1.58	0.33
50	A 50	3.40	0.22	0.86	0.09	84	A 84	-0.72	0.64	0.55	125	A125	1.47	20	-1.05	-0.34	0.98	0.05	1.07	0.36

****

TABLE CONTINUED

SEQ NUM	ITEM NAME	ITEM DIFF	STD ERROR	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	INFO AMT	FIT T-TESTS BETWN	FIT T-TESTS TOTAL	WTD MNSQ	MNSQ SD	DISC INDX	POINT BISER
51	A 51	0.11	0.15	0.33	2.03	74	A 74	-0.65	0.60	0.47	22	A 22	0.41	45	-0.94	-0.33	0.98	0.05	1.04	0.36
52	A 52	-1.83	0.26	1.61	-0.38	120	A120	-0.62	1.25	-0.53	16	A 16	-1.58	14	-0.51	-0.32	0.93	0.18	1.23	0.27
53	A 53	-3.07	0.45	1.34	0.06	4	A 4	-0.62	0.20	0.98	11	A 11	-1.98	4	1.31	-0.29	0.92	0.23	1.78	0.31
54	A 54	-0.72	0.18	1.72	-1.55	36	A 36	-0.56	1.42	-0.75	41	A 41	3.67	31	-0.62	-0.28	0.94	0.19	1.20	0.26
55	A 55	0.82	0.14	0.56	1.99	71	A 71	-0.45	0.57	1.01	82	A 82	-1.47	51	0.55	-0.26	0.95	0.17	1.54	0.32
56	A 56	-2.72	0.38	1.17	0.06	12	A 12	-0.42	1.08	0.19	68	A 68	-1.70	6	1.11	-0.26	0.94	0.20	1.49	0.29
57	A 57	0.95	0.14	1.26	-1.38	23	A 23	-0.40	1.09	-0.13	101	A101	-0.02	52	-0.93	-0.23	0.98	0.07	1.11	0.34
58	A 58	-2.88	0.41	1.52	0.03	79	A 79	-0.29	1.30	-1.08	3	A 3	-1.53	5	0.61	-0.22	0.95	0.18	1.37	0.29
59	A 59	3.35	0.22	0.80	0.34	95	A 95	-0.27	1.18	-0.79	10	A 10	-1.70	21	0.86	-0.21	0.95	0.20	1.47	0.28
60	A 60	0.62	0.14	1.35	-1.54	83	A 83	-0.27	1.02	-0.06	44	A 44	-1.83	50	0.02	-0.21	0.94	0.21	1.50	0.29
61	A 61	-0.17	0.16	1.79	-2.00	80	A 80	-0.27	1.29	-1.03	96	A 96	-2.58	41	0.60	-0.16	0.92	0.32	1.59	0.26
62	A 62	-2.34	0.32	0.97	-0.05	42	A 42	-0.27	1.45	-1.05	64	A 64	-1.37	9	0.48	-0.16	0.97	0.16	1.41	0.29
63	A 63	1.18	0.14	1.38	-2.14	75	A 75	-0.24	0.85	0.38	23	A 23	-0.40	51	-0.92	-0.13	0.99	0.09	1.09	0.30
64	A 64	-1.37	0.22	1.41	-0.16	113	A113	-0.19	0.79	0.47	18	A 18	0.22	20	-0.42	-0.10	0.99	0.06	1.02	0.33
65	A 65	2.39	0.16	0.37	1.47	81	A 81	-0.19	1.23	-0.82	98	A 98	-2.72	37	0.22	-0.10	0.93	0.35	1.76	0.25
66	A 66	-1.32	0.22	1.69	-0.82	61	A 61	-0.17	1.79	-2.00	78	A 78	-0.88	21	-0.24	-0.09	0.98	0.12	0.89	0.26
67	A 67	-0.95	0.19	0.81	0.41	119	A119	-0.14	1.73	-1.88	20	A 20	-1.32	27	0.75	-0.08	0.98	0.16	1.28	0.27
68	A 68	-1.70	0.25	1.49	-0.26	101	A101	-0.02	1.11	-0.23	15	A 15	3.06	16	-0.91	-0.08	0.98	0.13	1.07	0.26
69	A 69	0.55	0.14	1.04	0.26	106	A106	0.04	1.52	-1.72	24	A 24	-2.34	50	1.25	-0.07	0.95	0.28	1.77	0.24
70	A 70	0.95	0.14	1.48	-2.53	105	A105	0.07	0.67	0.97	49	A 49	-1.98	52	-1.49	-0.07	0.97	0.23	1.22	0.21
71	A 71	-0.45	0.17	0.57	1.01	51	A 51	0.11	0.33	2.03	72	A 72	-1.90	36	0.66	-0.06	0.97	0.22	1.10	0.20
72	A 72	-1.90	0.27	1.10	-0.06	31	A 31	0.11	0.85	0.76	83	A 83	-0.27	13	0.81	-0.06	0.99	0.08	1.02	0.29
73	A 73	0.18	0.15	1.59	-2.22	115	A115	0.15	1.21	-0.88	62	A 62	-2.34	46	-0.79	-0.05	0.96	0.28	0.97	0.17
74	A 74	-0.65	0.18	0.60	0.47	73	A 73	0.18	1.59	-2.22	93	A 93	-2.88	32	0.03	-0.05	0.94	0.38	1.54	0.22
75	A 75	-0.24	0.16	0.85	0.38	18	A 18	0.22	1.02	-0.10	6	A 6	-0.85	39	1.64	-0.05	0.99	0.12	1.22	0.29
76	A 76	3.06	0.26	0.25	0.38	1	A 1	0.24	1.13	-0.84	7	A 7	-2.34	14	-0.85	-0.03	0.96	0.28	1.14	0.18
77	A 77	3.06	0.20	-0.17	1.20	40	A 40	0.24	1.53	-2.19	137	A137	2.81	25	2.16	-0.03	0.99	0.11	0.75	0.20
78	A 78	-0.88	0.19	0.89	-0.09	43	A 43	0.32	0.24	2.83	90	A 90	-2.72	28	-0.73	-0.02	0.95	0.35	0.73	0.12
79	A 79	-0.29	0.16	1.30	-1.08	8	A 8	0.39	0.78	0.97	32	A 32	-2.24	39	-0.55	-0.01	0.97	0.27	1 20	0.19
80	A 80	-0.27	0.16	1.29	-1.03	22	A 22	0.41	1.04	-0.33	92	A 92	-2.24	39	1.11	0.00	0.98	0.27	0.82	0.12
81	A 81	-0.19	0.16	1.23	-0.82	13	A 13	0.45	1.83	-3.75	89	A 89	-2.88	40	2.13	0.01	0.96	0.38	0.30	0.07
82	A 82	-1.47	0.23	1.54	-0.26	108	A108	0.47	1.11	-0.69	129	A129	2.71	19	-1.06	0.03	1.00	0.11	0.95	0.26
83	A 83	-0.27	0.16	1.02	-0.06	126	A126	0.47	0.85	0.04	58	A 58	-2.88	39	0.30	0.03	0.97	0.38	1.52	0.17
84	A 84	-0.72	0.18	0.64	0.55	27	A 27	0.53	1.47	-2.30	126	A126	0.47	31	-0.71	0.04	1.00	0.05	0.85	0.33
85	A 85	-1.53	0.23	0.60	0.28	128	A128	0.53	0.94	0.36	132	A132	1.64	18	-1.20	0.05	1.00	0.05	0.97	0.32
86	A 86	-1.32	0.22	0.16	0.51	69	A 69	0.55	1.04	0.26	56	A 56	-2.72	21	-0.17	0.06	0.98	0.35	1.17	0.14
87	A 87	-0.85	0.19	0.14	1.03	117	A117	0.61	0.12	3.61	130	A130	0.86	29	-0.30	0.06	1.00	0.04	1.00	0.33
88	A 88	-2.06	0.29	0.63	0.09	60	A 60	0.62	1.35	-1.54	53	A 53	-3.07	12	-1.14	0.06	0.97	0.42	1.34	0.15
89	A 89	-2.88	0.41	0.30	0.01	138	A138	0.64	1.22	-0.96	88	A 88	-2.06	5	0.44	0.09	1.00	0.24	0.63	0.09
90	A 90	-2.72	0.38	0.73	-0.02	26	A 26	0.68	1.22	-0.96	50	A 50	3.40	6	-0.33	0.09	1.01	0.16	0.86	0.18
91	A 91	-2.72	0.38	0.89	0.11	136	A136	0.72	1.29	-1.68	45	A 45	-1.77	6	-0.64	0.10	1.01	0.20	0.86	0.15
92	A 92	-2.24	0.31	0.82	0.00	48	A 48	0.76	1.05	0.17	94	A 94	-4.00	10	-0.48	0.11	0.92	0.68	1.93	0.21
93	A 93	-2.88	0.41	1.54	-0.05	110	A110	0.78	0.04	4.22	91	A 91	-2.72	5	0.02	0.11	1.00	0.35	0.89	0.06
94	A 94	-4.00	0.70	1.93	0.11	123	A123	0.78	0.19	3.53	21	A 21	-3.59	2	1.12	0.13	0.97	0.55	0.79	0.07
95	A 95	-0.27	0.16	1.18	-0.79	55	A 55	0.82	0.56	1.99	48	A 48	0.76	39	-0.14	0.17	1.01	0.04	1.05	0.32
96	A 96	-2.58	0.36	1.59	-0.16	97	A 97	0.86	0.92	0.32	12	A 12	-0.42	7	1.61	0.19	1.01	0.09	1.08	0.29
97	A 97	0.86	0.14	0.92	0.32	130	A130	0.86	1.00	0.06	69	A 69	0.55	51	1.10	0.26	1.01	0.05	1.04	0.32
98	A 98	-2.72	0.38	1.76	-0.10	57	A 57	0.95	1.26	-1.38	85	A 85	-1.53	6	-0.61	0.28	1.04	0.18	0.60	0.10
99	A 99	1.85	0.15	-0.34	4.03	70	A 70	0.95	1.48	-2.53	97	A 97	0.86	45	-0.63	0.32	1.01	0.04	0.92	0.32
100	A100	-1.07	0.20	1.54	-0.73	17	A 17	1.01	1.69	-3.85	59	A 59	3.35	25	1.59	0.34	1.05	0.16	0.80	0.14
101	A101	-0.02	0.15	1.11	-0.23	63	A 63	1.18	1.38	-2.14	128	A128	0.53	43	0.05	0.36	1.02	0.05	0.94	0.31

TABLE CONTINUED

SEQ NUM	ITEM NAME	ITEM DIFF	STD ERROR	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	INFO AMT	FIT T-TESTS BETWN	FIT T-TESTS TOTAL	WTD MNSQ MNSQ	WTD MNSQ SD	DISC POINT INDX	DISC POINT BISER
102	A102	1.55	0.14	0.70	0.91	127	A127	1.22	1.14	-0.67	104	A104	-0.75	49	0.56	0.36	1.04	0.11	0.58	0.17
103	A103	1.70	0.14	0.35	2.54	25	A 25	1.28	0.82	0.81	140	A140	2.39	47	0.59	0.37	1.03	0.09	0.85	0.24
104	A104	-0.75	0.18	0.58	0.36	37	A 37	1.34	0.54	1.83	76	A 76	3.86	31	1.79	0.38	1.07	0.21	0.25	0.00
105	A105	0.07	0.15	0.67	0.97	46	A 46	1.36	1.59	-2.84	75	A 75	-0.24	44	-1.90	0.38	1.03	0.08	0.85	0.26
106	A106	0.04	0.15	1.52	-1.72	30	A 30	1.40	0.50	2.37	67	A 67	-0.95	44	0.75	0.41	1.05	0.12	0.81	0.18
107	A107	1.68	0.14	1.10	-0.79	112	A112	1.42	0.18	3.78	113	A113	-0.19	48	0.42	0.47	1.03	0.08	0.79	0.26
108	A108	0.47	0.14	1.11	-0.69	14	A 14	1.44	0.58	2.15	74	A 74	-0.65	49	0.98	0.47	1.05	0.10	0.60	0.17
109	A109	2.98	0.19	0.64	0.57	5	A 5	1.44	0.67	1.61	86	A 86	-1.32	27	2.90	0.51	1.07	0.16	0.16	0.03
110	A110	0.78	0.14	0.04	4.22	124	A124	1.45	1.21	-0.92	84	A 84	-0.72	51	0.64	0.55	1.06	0.11	0.64	0.16
111	A111	1.53	0.14	0.30	2.84	125	A125	1.47	1.07	-0.34	109	A109	2.98	49	0.98	0.57	1.07	0.13	0.64	0.15
112	A112	1.42	0.14	0.18	3.78	39	A 39	1.49	1.21	-0.80	135	A135	2.56	50	2.37	0.59	1.05	0.10	0.54	0.14
113	A113	-0.19	0.16	0.79	0.47	114	A114	1.51	1.68	-3.55	131	A131	2.07	40	-0.99	0.75	1.05	0.07	0.78	0.23
114	A114	1.51	0.14	1.68	-3.55	111	A111	1.53	0.30	2.84	31	A 31	0.11	49	0.38	0.76	1.05	0.06	0.85	0.25
115	A115	0.15	0.15	1.21	-0.88	102	A102	1.55	0.70	0.91	25	A 25	1.28	45	0.13	0.81	1.04	0.05	0.82	0.29
116	A116	1.66	0.14	1.56	-2.45	47	A 47	1.62	1.14	-0.60	102	A102	1.55	48	1.58	0.91	1.05	0.05	0.70	0.26
117	A117	0.61	0.14	0.12	3.61	139	A139	1.62	1.26	-1.05	8	A 8	0.39	50	0.51	0.97	1.05	0.05	0.78	0.27
118	A118	-1.28	0.21	1.26	-0.40	132	A132	1.64	0.97	0.05	105	A105	0.07	22	0.56	0.97	1.06	0.06	0.67	0.22
119	A119	-0.14	0.16	1.73	-1.88	116	A116	1.66	1.56	-2.45	4	A 4	-0.62	41	4.15	0.98	1.10	0.10	0.20	0.09
120	A120	-0.62	0.17	1.25	-0.53	107	A107	1.68	1.10	-0.79	71	A 71	-0.45	33	0.51	1.01	1.09	0.09	0.57	0.14
121	A121	1.74	0.15	1.75	-3.10	103	A103	1.70	0.35	2.54	87	A 87	-0.85	47	3.03	1.03	1.12	0.12	0.14	0.03
122	A122	-1.32	0.22	1.42	-0.40	121	A121	1.74	1.75	-3.10	133	A133	2.29	21	2.09	1.03	1.08	0.08	0.47	0.15
123	A123	0.78	0.14	0.19	3.53	99	A 99	1.85	-0.34	4.03	77	A 77	3.06	51	5.39	1.20	1.16	0.13	-0.17	-0.05
124	A124	1.45	0.14	1.21	-0.92	38	A 38	1.89	0.29	1.68	65	A 65	2.39	50	2.29	1.47	1.13	0.09	0.37	0.09
125	A125	1.47	0.14	1.07	-0.34	134	A134	2.05	0.29	1.83	5	A 5	1.44	50	2.10	1.61	1.08	0.05	0.67	0.22
126	A126	0.47	0.14	0.85	0.04	131	A131	2.07	0.78	0.75	38	A 38	1.89	49	3.93	1.68	1.11	0.06	0.29	0.13
127	A127	1.22	0.14	1.14	-0.67	34	A 34	2.09	1.30	-0.84	37	A 37	1.34	51	4.03	1.83	1.09	0.05	0.54	0.19
128	A128	0.53	0.14	0.94	0.36	133	A133	2.29	0.47	1.03	134	A134	2.05	50	2.43	1.83	1.13	0.07	0.29	0.10
129	A129	2.71	0.18	0.95	0.03	65	A 65	2.39	0.37	1.47	55	A 55	0.82	31	2.08	1.99	1.09	0.04	0.56	0.22
130	A130	0.86	0.14	1.00	0.06	140	A140	2.39	0.85	0.37	51	A 51	0.11	51	3.02	2.03	1.13	0.06	0.33	0.12
131	A131	2.07	0.15	0.78	0.75	135	A135	2.56	0.54	0.59	14	A 14	1.44	42	1.40	2.15	1.11	0.05	0.58	0.21
132	A132	1.64	0.14	0.97	0.05	129	A129	2.71	0.95	0.03	30	A 30	1.40	48	2.44	2.37	1.12	0.05	0.50	0.18
133	A133	2.29	0.16	0.47	1.03	137	A137	2.81	0.75	-0.03	103	A103	1.70	39	2.08	2.54	1.15	0.06	0.35	0.15
134	A134	2.05	0.15	0.29	1.83	109	A109	2.98	0.64	0.57	43	A 43	0.32	43	2.81	2.83	1.16	0.05	0.24	0.11
135	A135	2.56	0.17	0.54	0.59	77	A 77	3.06	-0.17	1.20	111	A111	1.53	34	2.39	2.84	1.15	0.05	0.30	0.12
136	A136	0.72	0.14	1.29	-1.68	15	A 15	3.06	1.07	-0.08	123	A123	0.78	51	3.63	3.53	1.17	0.04	0.19	0.13
137	A137	2.81	0.18	0.75	-0.03	59	A 59	3.35	0.80	0.34	117	A117	0.61	30	3.36	3.61	1.18	0.05	0.12	0.10
138	A138	0.64	0.14	1.22	-0.96	50	A 50	3.40	0.86	0.09	112	A112	1.42	50	2.97	3.78	1.19	0.05	0.18	0.10
139	A139	1.62	0.14	1.26	-1.05	41	A 41	3.67	1.20	-0.28	99	A 99	1.85	48	5.46	4.03	1.27	0.06	-0.34	-0.04
140	A140	2.39	0.16	0.85	0.37	76	A 76	3.86	0.25	0.38	110	A110	0.78	37	4.08	4.22	1.20	0.04	0.04	0.08
	MEAN	0.00		1.05	-0.05										1.00	-0.05	0.99	0.13		
	S.D.	1.68		0.48	1.40										1.41	1.40	0.09	0.11		

140 ITEMS CALIBRATED ON 236 PERSONS
236 MEASURABLE PERSONS WITH MEAN ABILITY = 1.01 AND STD. DEV. = 0.73

SEQ NUM	ITEM NAME	ITEM DIFF	STD ERROR	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	INFO AMT	FIT T-TESTS BETWN	FIT T-TESTS TOTAL	WTD MNSQ	MNSQ SD	DISC INDX	POINT BISER
1	B 1	-0.01	0.15	1.80	-4.60	44	B 44	-3.76	1.22	0.06	34	B 34	0.61	45	3.54	-5.13	0.76	0.05	1.85	0.64
2	B 2	1.51	0.16	1.12	-0.58	78	B 78	-3.76	1.22	0.10	101	B101	1.08	39	3.98	-4.97	0.74	0.06	1.95	0.65
3	B 3	0.48	0.15	1.32	-1.95	38	B 38	-3.34	1.65	-0.07	1	B 1	-0.01	47	3.13	-4.60	0.77	0.05	1.80	0.62
4	B 4	-0.62	0.16	0.83	0.72	65	B 65	-2.70	1.44	-0.07	23	B 23	0.35	38	3.08	-4.48	0.79	0.05	1.82	0.61
5	B 5	-0.83	0.17	0.94	0.24	119	B119	-2.61	0.80	0.11	122	B122	0.46	35	2.98	-4.30	0.80	0.05	1.79	0.60
6	B 6	1.15	0.15	1.65	-3.47	42	B 42	-2.22	1.31	-0.30	6	B 6	1.15	43	3.73	-3.47	0.80	0.06	1.65	0.57
7	B 7	-1.69	0.21	1.39	-0.44	40	B 40	-2.04	1.59	-0.48	105	B105	0.46	22	2.71	-3.40	0.84	0.05	1.60	0.55
8	B 8	0.76	0.15	0.78	1.36	50	B 50	-1.93	1.73	-0.78	43	B 43	-0.23	46	2.60	-3.34	0.82	0.06	1.76	0.57
9	B 9	-0.30	0.15	1.52	-2.00	100	B100	-1.73	1.21	-0.30	29	B 29	0.06	42	2.29	-3.20	0.84	0.05	1.61	0.55
10	B 10	-0.64	0.16	1.76	-2.33	19	B 19	-1.73	1.45	-0.56	118	B118	0.91	38	2.07	-3.14	0.84	0.05	1.61	0.55
11	B 11	-1.52	0.20	1.52	-0.82	7	B 7	-1.69	1.39	-0.44	120	B120	0.18	25	2.05	-3.05	0.85	0.05	1.56	0.54
12	B 12	-0.94	0.17	1.11	-0.48	36	B 36	-1.69	0.83	0.09	96	B 96	0.48	34	2.14	-3.04	0.86	0.05	1.55	0.53
13	B 13	0.10	0.15	1.46	-2.30	72	B 72	-1.69	1.08	-0.11	71	B 71	1.74	46	4.75	-2.69	0.79	0.08	1.54	0.51
14	B 14	0.44	0.15	-0.16	5.40	26	B 26	-1.60	1.57	-0.90	25	B 25	0.78	47	1.97	-2.69	0.86	0.05	1.46	0.51
15	B 15	2.65	0.21	0.46	0.91	41	B 41	-1.56	1.61	-0.70	135	B135	1.04	22	0.99	-2.64	0.86	0.06	1.52	0.52
16	B 16	-1.16	0.18	0.48	0.87	11	B 11	-1.52	1.52	-0.82	123	B123	-0.01	30	1.77	-2.46	0.87	0.05	1.42	0.50
17	B 17	-0.78	0.17	1.48	-1.54	69	B 69	-1.40	1.78	-1.47	64	B 64	-0.49	36	2.55	-2.34	0.85	0.07	1.60	0.52
18	B 18	-0.59	0.16	1.41	-1.43	45	B 45	-1.37	1.20	-0.12	10	B 10	-0.64	39	2.52	-2.33	0.84	0.07	1.76	0.53
19	B 19	-1.73	0.21	1.45	-0.56	103	B103	-1.26	0.93	0.04	13	B 13	0.10	21	1.01	-2.30	0.88	0.05	1.46	0.49
20	B 20	-1.13	0.18	1.74	-1.71	31	B 31	-1.26	1.86	-1.76	138	B138	1.53	31	3.74	-2.21	0.85	0.07	1.46	0.47
21	B 21	0.61	0.15	0.96	0.11	70	B 70	-1.19	1.62	-1.42	37	B 37	-0.01	47	2.00	-2.07	0.89	0.05	1.41	0.48
22	B 22	-1.19	0.18	1.80	-1.75	22	B 22	-1.19	1.80	-1.75	39	B 39	0.08	30	1.31	-2.00	0.90	0.05	1.36	0.48
23	B 23	0.35	0.15	1.82	-4.48	35	B 35	-1.19	1.10	-0.31	9	B 9	-0.30	47	1.30	-2.00	0.88	0.06	1.52	0.49
24	B 24	-0.97	0.17	0.57	1.05	66	B 66	-1.16	1.73	-1.49	3	B 3	0.48	33	1.29	-1.95	0.91	0.05	1.32	0.48
25	B 25	0.78	0.15	1.46	-2.69	16	B 16	-1.16	0.48	0.87	46	B 46	-0.03	46	1.57	-1.93	0.90	0.05	1.33	0.47
26	B 26	-1.60	0.21	1.57	-0.90	84	B 84	-1.13	1.08	-0.42	28	B 28	-0.83	23	2.59	-1.90	0.85	0.08	1.72	0.51
27	B 27	0.48	0.15	0.85	0.72	20	B 20	-1.13	1.74	-1.71	62	B 62	-0.72	47	2.18	-1.90	0.86	0.08	1.56	0.49
28	B 28	-0.83	0.17	1.72	-1.90	49	B 49	-1.06	1.00	-0.09	31	B 31	-1.26	35	3.10	-1.76	0.82	0.11	1.86	0.52
29	B 29	0.06	0.15	1.61	-3.20	79	B 79	-1.06	1.09	-0.08	22	B 22	-1.19	45	2.94	-1.75	0.83	0.10	1.80	0.51
30	B 30	-0.97	0.17	1.38	-0.99	80	B 80	-1.03	1.09	-0.06	20	B 20	-1.13	33	2.53	-1.71	0.84	0.10	1.74	0.51
31	B 31	-1.26	0.19	1.86	-1.76	75	B 75	-1.00	1.16	-0.31	48	B 48	-0.94	29	2.29	-1.71	0.85	0.09	1.62	0.49
32	B 32	1.69	0.16	0.44	1.63	73	B 73	-1.00	1.50	-0.91	91	B 91	-0.57	37	2.42	-1.64	0.89	0.07	1.44	0.46
33	B 33	-0.44	0.16	0.81	0.54	30	B 30	-0.97	1.38	-0.99	116	B116	0.42	41	0.63	-1.56	0.92	0.05	1.39	0.46
34	B 34	0.61	0.15	1.85	-5.13	24	B 24	-0.97	0.57	1.05	17	B 17	-0.78	47	1.24	-1.54	0.88	0.08	1.48	0.46
35	B 35	-1.19	0.18	1.10	-0.31	12	B 12	-0.94	1.11	-0.48	66	B 66	-1.16	30	2.60	-1.49	0.85	0.10	1.73	0.47
36	B 36	-1.69	0.21	0.83	0.09	48	B 48	-0.94	1.62	-1.71	69	B 69	-1.40	22	3.11	-1.47	0.83	0.12	1.78	0.49
37	B 37	-0.01	0.15	1 41	-2.07	104	B104	-0.86	1.46	-1.13	83	B 83	-0.30	45	1.15	-1.46	0.91	0.06	1.29	0.44
38	B 38	-3.34	0.41	1.65	-0.07	5	B 5	-0.83	0.94	0.24	51	B 51	1.11	5	0.81	-1.45	0.92	0.06	1.31	0.45
39	B 39	0.08	0.15	1.36	-2.00	28	B 28	-0.83	1.72	-1.90	18	B 18	-0.59	46	2.04	-1.43	0.90	0.07	1.41	0.46
40	B 40	-2.04	0.24	1.59	-0.48	99	B 99	-0.80	1.52	-1.20	70	B 70	-1.19	17	2.38	-1.42	0.86	0.10	1.62	0.46
41	B 41	-1.56	0.20	1.61	-0.70	17	B 17	-0.78	1.48	-1.54	58	B 58	-0.72	24	0.95	-1.31	0.90	0.08	1.49	0.45
42	B 42	-2.22	0.26	1.31	-0.30	58	B 58	-0.72	1.49	-1.31	81	B 81	-0.42	15	0.17	-1.24	0.92	0.06	1.32	0.43
43	B 43	-0.23	0.15	1.76	-3.34	62	B 62	-0.72	1.56	-1.90	99	B 99	-0.80	43	1.63	-1.20	0.90	0.08	1.52	0.45
44	B 44	-3.76	0.50	1.22	0.06	10	B 10	-0.64	1.76	-2.33	104	B104	-0.86	3	1.20	-1.13	0.91	0.08	1.46	0.43
45	B 45	-1.37	0.19	1.20	-0.12	82	B 82	-0.64	1.05	-0.13	90	B 90	-0.23	27	0.68	-1.06	0.94	0.06	1.26	0.42
46	B 46	-0.03	0.15	1.33	-1.93	4	B 4	-0.62	0.83	0.72	30	B 30	-0.97	45	0.80	-0.99	0.91	0.09	1.38	0.42
47	B 47	1.36	0.16	0.79	-0.12	92	B 92	-0.59	1.31	-0.94	92	B 92	-0.59	41	0.49	-0.94	0.93	0.07	1.31	0.40
48	B 48	-0.94	0.17	1.62	-1.71	18	B 18	-0.59	1.41	-1.43	54	B 54	0.21	34	1.78	-0.93	0.95	0.05	1.21	0.42
49	B 49	-1.06	0.18	1.00	-0.09	91	B 91	-0.57	1.44	-1.64	73	B 73	-1.00	32	2.22	-0.91	0.92	0.09	1.50	0.43
50	B 50	-1.93	0.23	1.73	-0.78	64	B 64	-0.49	1.60	-2.34	26	B 26	-1.60	18	1.46	-0.90	0.88	0.14	1.57	0.42

TABLE CONTINUED

SEQ NUM	ITEM NAME	ITEM DIFF	STD ERROR	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	INFO AMT	FIT T-TESTS BETWN	FIT T-TESTS TOTAL	WTD MNSQ MNSQ	WTD MNSQ SD	DISC POINT INDX	DISC POINT BISER
51	B 51	1.11	0.15	1.31	-1.45	68	B 68	-0.44	0.99	-0.21	130	B130	0.42	44	0.19	-0.83	0.96	0.05	1.16	0.41
52	B 52	-0.05	0.15	0.44	2.10	33	B 33	-0.44	0.81	0.54	11	B 11	-1.52	45	1.03	-0.82	0.89	0.13	1.52	0.40
53	B 53	0.86	0.15	0.96	0.45	81	B 81	-0.42	1.32	-1.24	50	B 50	-1.93	46	2.17	-0.78	0.87	0.17	1.73	0.42
54	B 54	0.21	0.15	1.21	-0.93	98	B 98	-0.32	0.90	0.49	117	B117	1.53	46	-0.96	-0.77	0.94	0.07	1.25	0.41
55	B 55	-0.10	0.15	1.17	-0.62	86	B 86	-0.30	0.85	0.72	107	B107	0.89	44	0.09	-0.76	0.96	0.05	1.15	0.40
56	B 56	1.06	0.15	0.59	1.76	83	B 83	-0.30	1.29	-1.46	41	B 41	-1.56	44	1.71	-0.70	0.90	0.13	1.61	0.40
57	B 57	0.16	0.15	0.98	0.45	114	B114	-0.30	0.44	2.64	132	B132	0.29	46	1.29	-0.70	0.96	0.05	1.17	0.41
58	B 58	-0.72	0.16	1.49	-1.31	9	B 9	-0.30	1.52	-2.00	55	B 55	-0.10	37	1.87	-0.62	0.96	0.06	1.17	0.39
59	B 59	0.01	0.15	0.05	4.50	74	B 74	-0.23	0.90	0.70	2	B 2	1.51	45	-0.22	-0.58	0.96	0.07	1.12	0.38
60	B 60	4.79	0.51	-0.13	0.20	43	B 43	-0.23	1.76	-3.34	19	B 19	-1.73	3	1.58	-0.56	0.91	0.15	1.45	0.36
61	B 61	0.76	0.15	0.75	1.58	90	B 90	-0.23	1.26	-1.06	12	B 12	-0.94	46	0.00	-0.48	0.96	0.09	1.11	0.34
62	B 62	-0.72	0.16	1.56	-1.90	55	B 55	-0.10	1.17	-0.62	40	B 40	-2.04	37	0.76	-0.48	0.91	0.18	1.59	0.36
63	B 63	2.79	0.22	0.30	0.54	52	B 52	-0.05	0.44	2.10	7	B 7	-1.69	20	0.53	-0.44	0.93	0.14	1.39	0.35
64	B 64	-0.49	0.16	1.60	-2.34	76	B 76	-0.05	0.38	2.88	84	B 84	-1.13	40	0.06	-0.42	0.96	0.10	1.08	0.32
65	B 65	-2.70	0.31	1.44	-0.07	46	B 46	-0.03	1.33	-1.93	35	B 35	-1.19	10	-0.31	-0.31	0.96	0.10	1.10	0.32
66	B 66	-1.16	0.18	1.73	-1.49	37	B 37	-0.01	1.41	-2.07	75	B 75	-1.00	30	-0.03	-0.31	0.97	0.09	1.16	0.33
67	B 67	0.21	0.15	0.70	1.69	87	B 87	-0.01	0.20	3.47	100	B100	-1.73	46	-0.63	-0.30	0.95	0.15	1.21	0.30
68	B 68	-0.44	0.16	0.99	-0.21	1	B 1	-0.01	1.80	-4.60	42	B 42	-2.22	41	-0.71	-0.30	0.93	0.20	1.31	0.29
69	B 69	-1.40	0.19	1.78	-1.47	123	B123	-0.01	1.42	-2.46	68	B 68	-0.44	26	0.92	-0.21	0.98	0.07	0.99	0.34
70	B 70	-1.19	0.18	1.62	-1.42	88	B 88	0.01	0.27	3.13	82	B 82	-0.64	30	-0.99	-0.13	0.99	0.07	1.05	0.33
71	B 71	1.74	0.17	1.54	-2.69	59	B 59	0.01	0.05	4.50	45	B 45	-1.37	36	0.28	-0.12	0.98	0.12	1.20	0.30
72	B 72	-1.69	0.21	1.08	-0.11	29	B 29	0.06	1.61	-3.20	47	B 47	1.36	22	6.79	-0.12	0.99	0.07	0.79	0.27
73	B 73	-1.00	0.17	1.50	-0.91	125	B125	0.06	-0.47	6.56	72	B 72	-1.69	33	-0.71	-0.11	0.98	0.14	1.08	0.27
74	B 74	-0.23	0.15	0.90	0.70	39	B 39	0.08	1.36	-2.00	133	B133	1.31	43	-1.10	-0.10	0.99	0.06	1.03	0.36
75	B 75	-1.00	0.17	1.16	-0.31	106	B106	0.08	0.81	0.80	49	B 49	-1.06	33	-2.48	-0.09	0.99	0.10	1.00	0.30
76	B 76	-0.05	0.15	0.38	2.88	13	B 13	0.10	1.46	-2.30	79	B 79	-1.06	45	-1.28	-0.08	0.99	0.10	1.09	0.30
77	B 77	1.24	0.15	0.19	3.16	109	B109	0.12	0.48	2.96	38	B 38	-3.34	42	0.61	-0.07	0.93	0.38	1.65	0.24
78	B 78	-3.76	0.50	1.22	0.10	97	B 97	0.14	0.17	4.20	65	B 65	-2.70	3	-0.36	-0.07	0.96	0.27	1.44	0.24
79	B 79	-1.06	0.18	1.09	-0.08	57	B 57	0.16	0.98	0.45	80	B 80	-1.03	32	-1.01	-0.06	0.99	0.09	1.09	0.31
80	B 80	-1.03	0.18	1.09	-0.06	120	B120	0.18	1.56	-3.05	103	B103	-1.26	32	1.78	0.04	1.00	0.11	0.93	0.25
81	B 81	-0.42	0.16	1.32	-1.24	67	B 67	0.21	0.70	1.69	44	B 44	-3.76	41	-0.91	0.06	0.95	0.47	1.22	0.15
82	B 82	-0.64	0.16	1.05	-0.13	54	B 54	0.21	1.21	-0.93	36	B 36	-1.69	38	-0.48	0.09	1.01	0.14	0.83	0.20
83	B 83	-0.30	0.15	1.29	-1.46	102	B102	0.21	0.81	0.97	78	B 78	-3.76	42	0.10	0.10	0.97	0.47	1.22	0.11
84	B 84	-1.13	0.18	1.08	-0.42	132	B132	0.29	1.17	-0.70	21	B 21	0.61	31	0.59	0.11	1.00	0.05	0.96	0.37
85	B 85	0.82	0.15	-0.95	8.51	23	B 23	0.35	1.82	-4.48	119	B119	-2.61	46	-1.49	0.11	1.01	0.25	0.80	0.12
86	B 86	-0.30	0.15	0.85	0.72	95	B 95	0.40	0.67	1.62	60	B 60	4.79	42	2.97	0.20	1.02	0.47	-0.13	-0.07
87	B 87	-0.01	0.15	0.20	3.47	130	B130	0.42	1.16	-0.83	5	B 5	-0.83	45	0.74	0.24	1.02	0.08	0.94	0.28
88	B 88	0.01	0.15	0.27	3.13	116	B116	0.42	1.39	-1.56	128	B128	1.94	45	1.91	0.25	1.02	0.09	0.66	0.21
89	B 89	0.95	0.15	0.75	1.31	14	B 14	0.44	-0.16	5.40	134	B134	2.41	45	4.41	0.25	1.03	0.12	0.56	0.15
90	B 90	-0.23	0.15	1.26	-1.06	105	B105	0.46	1.60	-3.40	93	B 93	1.56	43	0.50	0.33	1.02	0.07	0.95	0.33
91	B 91	-0.57	0.16	1.44	-1.64	122	B122	0.46	1.79	-4.30	57	B 57	0.16	39	1.12	0.45	1.02	0.05	0.98	0.34
92	B 92	-0.59	0.16	1.31	-0.94	3	B 3	0.48	1.32	-1.95	53	B 53	0.86	39	0.14	0.45	1.02	0.05	0.96	0.35
93	B 93	1.56	0.16	0.95	0.33	96	B 96	0.48	1.55	-3.04	98	B 98	-0.32	39	-0.84	0.49	1.03	0.06	0.90	0.31
94	B 94	1.51	0.16	0.77	0.95	27	B 27	0.48	0.85	0.72	33	B 33	-0.44	39	-0.45	0.54	1.03	0.07	0.81	0.30
95	B 95	0.40	0.15	0.67	1.62	21	B 21	0.61	0.96	0.11	136	B136	1.97	47	-0.24	0.54	1.05	0.09	0.73	0.23
96	B 96	0.48	0.15	1.55	-3.04	34	B 34	0.61	1.85	-5.13	63	B 63	2.79	47	3.13	0.54	1.08	0.15	0.30	0.07
97	B 97	0.14	0.15	0.17	4.20	124	B124	0.67	-0.04	4.55	74	B 74	-0.23	46	0.30	0.70	1.04	0.06	0.90	0.30
98	B 98	-0.32	0.15	0.90	0.49	8	B 8	0.76	0.78	1.36	27	B 27	0.48	42	-0.78	0.72	1.04	0.05	0.85	0.33
99	B 99	-0.80	0.17	1.52	-1.20	61	B 61	0.76	0.75	1.58	4	B 4	-0.62	36	0.28	0.72	1.05	0.07	0.83	0.26
100	B100	-1.73	0.21	1.21	-0.30	115	B115	0.78	0.66	1.35	86	B 86	-0.30	21	-0.48	0.72	1.04	0.06	0.85	0.29
101	B101	1.08	0.15	1.95	-4.97	25	B 25	0.78	1.46	-2.69	112	B112	1.11	44	-1.11	0.74	1.04	0.06	0.87	0.31

****

TABLE CONTINUED

SEQ NUM	ITEM NAME	ITEM DIFF	STD ERROR	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	INFO AMT	FIT T-TESTS BETWN	FIT T-TESTS TOTAL	WTD MNSQ	WTD MNSQ SD	DISC INDX	POINT BISER
102	B102	0.21	0.15	0.81	0.97	85	B 85	0.82	-0.95	8.51	106	B106	0.08	46	1.34	0.80	1.04	0.05	0.81	0.31
103	B103	-1.26	0.19	0.93	0.04	53	B 53	0.86	0.96	0.45	16	B 16	-1.16	29	2.01	0.87	1.09	0.10	0.48	0.13
104	B104	-0.86	0.17	1.46	-1.13	107	B107	0.89	1.15	-0.76	15	B 15	2.65	35	1.43	0.91	1.13	0.14	0.46	0.10
105	B105	0.46	0.15	1.60	-3.40	118	B118	0.91	1.61	-3.14	94	B 94	1.51	47	0.33	0.95	1.07	0.07	0.77	0.27
106	B106	0.08	0.15	0.81	0.80	89	B 89	0.95	0.75	1.31	102	B102	0.21	46	0.58	0.97	1.05	0.05	0.81	0.31
107	B107	0.89	0.15	1.15	-0.76	135	B135	1.04	1.52	-2.64	137	B137	1.15	45	1.20	1.02	1.06	0.06	0.76	0.27
108	B108	1.36	0.16	0.25	2.39	56	B 56	1.06	0.59	1.76	24	B 24	-0.97	41	2.25	1.05	1.09	0.09	0.57	0.17
109	B109	0.12	0.15	0.48	2.96	101	B101	1.08	1.95	-4.97	89	B 89	0.95	46	1.71	1.31	1.07	0.06	0.75	0.29
110	B110	1.08	0.15	0.61	1.57	110	B110	1.08	0.61	1.57	115	B115	0.78	44	1.40	1.35	1.07	0.05	0.66	0.28
111	B111	1.43	0.16	-0.35	4.79	126	B126	1.08	0.33	2.67	8	B 8	0.76	40	0.00	1.36	1.07	0.05	0.78	0.29
112	B112	1.11	0.15	0.87	0.74	113	B113	1.11	0.45	2.54	110	B110	1.08	44	0.18	1.57	1.09	0.06	0.61	0.25
113	B113	1.11	0.15	0.45	2.54	131	B131	1.11	0.44	2.33	61	B 61	0.76	44	1.13	1.58	1.08	0.05	0.75	0.27
114	B114	-0.30	0.15	0.44	2.64	51	B 51	1.11	1.31	-1.45	95	B 95	0.40	42	0.59	1.62	1.08	0.05	0.67	0.27
115	B115	0.78	0.15	0.66	1.35	112	B112	1.11	0.87	0.74	32	B 32	1.69	46	1.91	1.63	1.13	0.08	0.44	0.16
116	B116	0.42	0.15	1.39	-1.56	6	B 6	1.15	1.65	-3.47	67	B 67	0.21	47	0.69	1.69	1.09	0.05	0.70	0.26
117	B117	1.53	0.16	1.25	-0.77	137	B137	1.15	0.76	1.02	139	B139	2.34	39	4.16	1.76	1.21	0.12	-0.05	-0.02
118	B118	0.91	0.15	1.61	-3.14	77	B 77	1.24	0.19	3.16	56	B 56	1.06	45	0.59	1.76	1.10	0.06	0.59	0.24
119	B119	-2.61	0.30	0.80	0.11	133	B133	1.31	1.03	-0.10	52	B 52	-0.05	11	3.43	2.10	1.12	0.05	0.44	0.21
120	B120	0.18	0.15	1.56	-3.05	108	B108	1.36	0.25	2.39	140	B140	1.86	46	3.28	2.13	1.20	0.09	0.15	0.06
121	B121	2.03	0.18	-0.46	3.26	47	B 47	1.36	0.79	-0.12	131	B131	1.11	32	2.89	2.33	1.14	0.06	0.44	0.20
122	B122	0.46	0.15	1.79	-4.30	111	B111	1.43	-0.35	4.79	108	B108	1.36	47	2.84	2.39	1.17	0.07	0.25	0.15
123	B123	-0.01	0.15	1.42	-2.46	127	B127	1.43	0.16	2.87	113	B113	1.11	45	2.02	2.54	1.16	0.06	0.45	0.18
124	B124	0.67	0.15	-0.04	4.55	94	B 94	1.51	0.77	0.95	114	B114	-0.30	46	2.66	2.64	1.17	0.06	0.44	0.15
125	B125	0.06	0.15	-0.47	6.56	2	B 2	1.51	1.12	-0.58	129	B129	1.74	45	3.26	2.66	1.23	0.08	0.17	0.07
126	B126	1.08	0.15	0.33	2.67	117	B117	1.53	1.25	-0.77	126	B126	1.08	44	3.27	2.67	1.16	0.06	0.33	0.15
127	B127	1.43	0.16	0.16	2.87	138	B138	1.53	1.46	-2.21	127	B127	1.43	40	3.37	2.87	1.21	0.07	0.16	0.10
128	B128	1.94	0.17	0.66	0.25	93	B 93	1.56	0.95	0.33	76	B 76	-0.05	33	1.84	2.88	1.16	0.05	0.38	0.15
129	B129	1.74	0.17	0.17	2.66	32	B 32	1.69	0.44	1.63	109	B109	0.12	36	3.85	2.96	1.16	0.05	0.48	0.18
130	B130	0.42	0.15	1.16	-0.83	129	B129	1.74	0.17	2.66	88	B 88	0.01	47	3.15	3.13	1.17	0.05	0.27	0.15
131	B131	1.11	0.15	0.44	2.33	71	B 71	1.74	1.54	-2.69	77	B 77	1.24	44	3.72	3.16	1.21	0.06	0.19	0.10
132	B132	0.29	0.15	1.17	-0.70	140	B140	1.86	0.15	2.13	121	B121	2.03	46	6.54	3.26	1.35	0.10	-0.46	-0.15
133	B133	1.31	0.15	1.03	-0.10	128	B128	1.94	0.66	0.25	87	B 87	-0.01	42	3.08	3.47	1.20	0.05	0.20	0.12
134	B134	2.41	0.19	0.56	0.25	136	B136	1.97	0.73	0.54	97	B 97	0.14	26	3.06	4.20	1.23	0.05	0.17	0.09
135	B135	1.04	0.15	1.52	-2.64	121	B121	2.03	-0.46	3.26	59	B 59	0.01	44	4.90	4.50	1.26	0.05	0.05	0.05
136	B136	1.97	0.17	0.73	0.54	139	B139	2.34	-0.05	1.76	124	B124	0.67	33	4.61	4.55	1.25	0.05	-0.04	0.08
137	B137	1.15	0.15	0.76	1.02	134	B134	2.41	0.56	0.25	111	B111	1.43	43	6.14	4.79	1.37	0.07	-0.35	-0.08
138	B138	1.53	0.16	1.46	-2.21	15	B 15	2.65	0.46	0.91	14	B 14	0.44	39	4.94	5.40	1.29	0.05	-0.16	0.03
139	B139	2.34	0.19	-0.05	1.76	63	B 63	2.79	0.30	0.54	125	B125	0.06	27	6.97	6.56	1.38	0.05	-0.47	-0.10
140	B140	1.86	0.17	0.15	2.13	60	B 60	4.79	-0.13	0.20	85	B 85	0.82	34	8.90	8.51	1.52	0.05	-0.95	-0.23
	MEAN	0.00		1.03	-0.08										1.68	-0.08	0.99	0.09		
	S.D.	1.30		0.58	2.22										1.84	2.22	0.14	0.07		

140 ITEMS CALIBRATED ON 219 PERSONS
219 MEASURABLE PERSONS WITH MEAN ABILITY = 0.50 AND STD. DEV. = 0.80

SEQ NUM	ITEM NAME	ITEM DIFF	STD ERROR	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	INFO AMT	FIT T-TESTS BETWN	FIT T-TESTS TOTAL	WTD MNSQ	MNSQ SD	DISC INDX	POINT BISER
1	C 1	0.37	0.14	0.62	2.07	94	C 94	-5.13	1.71	0.28	103	C103	0.84	52	4.36	-5.90	0.73	0.05	1.97	0.67
2	C 2	-1.01	0.17	1.33	-0.85	89	C 89	-4.02	0.83	0.11	62	C 62	0.41	34	3.86	-5.41	0.76	0.05	1.85	0.65
3	C 3	0.60	0.14	1.34	-2.28	93	C 93	-3.73	1.49	0.06	34	C 34	0.54	53	4.17	-5.37	0.76	0.05	1.86	0.64
4	C 4	-0.39	0.15	1.61	-2.65	90	C 90	-3.32	0.84	0.06	33	C 33	0.24	45	3.43	-4.23	0.80	0.05	1.68	0.60
5	C 5	-0.79	0.16	1.64	-1.81	92	C 92	-3.32	0.73	0.12	15	C 15	1.07	38	2.83	-4.10	0.80	0.05	1.66	0.59
6	C 6	0.20	0.14	1.61	-3.69	37	C 37	-3.02	1.59	-0.04	19	C 19	0.26	51	2.49	-3.80	0.82	0.05	1.69	0.58
7	C 7	-1.40	0.19	1.51	-0.88	91	C 91	-2.89	0.69	0.09	12	C 12	1.09	27	3.15	-3.69	0.82	0.05	1.60	0.57
8	C 8	1.23	0.14	0.94	0.73	88	C 88	-2.68	0.76	0.15	6	C 6	0.20	50	2.35	-3.69	0.82	0.05	1.61	0.57
9	C 9	1.09	0.14	0.42	3.08	70	C 70	-2.28	1.13	-0.14	127	C127	0.20	51	2.39	-3.66	0.83	0.05	1.63	0.57
10	C 10	0.84	0.14	1.37	-2.63	104	C104	-2.14	1.46	-0.32	56	C 56	-0.35	53	3.81	-3.47	0.79	0.06	1.78	0.59
11	C 11	-1.97	0.23	1.20	-0.18	21	C 21	-2.08	0.89	0.13	96	C 96	0.79	18	1.90	-3.47	0.84	0.05	1.59	0.56
12	C 12	1.09	0.14	1.60	-3.69	64	C 64	-1.97	1.72	-0.67	107	C107	0.79	51	2.86	-3.46	0.84	0.05	1.51	0.56
13	C 13	0.47	0.14	0.93	0.44	11	C 11	-1.97	1.20	-0.18	50	C 50	-0.09	53	3.30	-3.32	0.82	0.06	1.67	0.56
14	C 14	0.96	0.14	1.10	-0.70	82	C 82	-1.92	0.93	0.11	16	C 16	0.01	52	2.62	-3.00	0.85	0.05	1.59	0.55
15	C 15	1.07	0.14	1.66	-4.10	66	C 66	-1.82	1.18	-0.20	106	C106	1.13	51	2.20	-2.95	0.85	0.05	1.57	0.54
16	C 16	0.01	0.14	1.59	-3.00	18	C 18	-1.77	1.56	-0.63	57	C 57	1.30	50	1.68	-2.89	0.84	0.06	1.51	0.54
17	C 17	-0.87	0.17	1.27	-0.84	52	C 52	-1.77	1.80	-0.85	129	C129	1.77	36	2.79	-2.70	0.82	0.07	1.58	0.54
18	C 18	-1.77	0.22	1.56	-0.63	24	C 24	-1.72	1.03	-0.02	4	C 4	-0.39	21	2.73	-2.65	0.84	0.07	1.61	0.54
19	C 19	0.26	0.14	1.69	-3.80	102	C102	-1.64	1.45	-0.52	10	C 10	0.84	52	1.53	-2.63	0.87	0.05	1.37	0.52
20	C 20	1.79	0.15	0.99	-0.43	85	C 85	-1.55	0.51	0.54	22	C 22	1.55	43	2.02	-2.43	0.85	0.06	1.48	0.51
21	C 21	-2.08	0.24	0.89	0.13	23	C 23	-1.51	1.32	-0.39	3	C 3	0.60	17	0.76	-2.28	0.89	0.05	1.34	0.50
22	C 22	1.55	0.15	1.48	-2.43	48	C 48	-1.44	1.43	-0.73	40	C 40	1.30	47	1.40	-2.10	0.88	0.06	1.36	0.48
23	C 23	-1.51	0.20	1.32	-0.39	68	C 68	-1.44	1.41	-0.65	28	C 28	-0.72	25	2.59	-1.86	0.86	0.08	1.65	0.50
24	C 24	-1.72	0.21	1.03	-0.02	7	C 7	-1.40	1.51	-0.88	114	C114	0.34	22	-0.09	-1.84	0.91	0.05	1.27	0.48
25	C 25	1.30	0.14	0.78	0.65	86	C 86	-1.30	0.38	1.03	5	C 5	-0.79	49	2.23	-1.81	0.85	0.08	1.64	0.50
26	C 26	-0.62	0.16	0.88	0.34	100	C100	-1.26	1.05	-0.37	65	C 65	-0.17	41	0.92	-1.79	0.90	0.06	1.37	0.47
27	C 27	3.05	0.21	0.39	0.65	67	C 67	-1.17	1.74	-1.59	105	C105	1.01	22	0.93	-1.72	0.91	0.05	1.26	0.46
28	C 28	-0.72	0.16	1.65	-1.86	54	C 54	-1.10	1.40	-0.79	36	C 36	-1.07	39	2.62	-1.65	0.84	0.10	1.68	0.49
29	C 29	-0.96	0.17	1.43	-1.19	87	C 87	-1.10	0.09	1.67	61	C 61	-0.22	35	0.89	-1.60	0.91	0.06	1.33	0.45
30	C 30	0.90	0.14	-0.24	6.13	36	C 36	-1.07	1.68	-1.65	35	C 35	-1.04	52	2.71	-1.59	0.85	0.10	1.67	0.48
31	C 31	1.96	0.16	1.04	-0.34	35	C 35	-1.04	1.67	-1.59	67	C 67	-1.17	41	3.03	-1.59	0.84	0.11	1.74	0.49
32	C 32	-0.41	0.15	1.37	-1.45	2	C 2	-1.01	1.33	-0.85	32	C 32	-0.41	44	0.97	-1.45	0.91	0.07	1.37	0.45
33	C 33	0.24	0.14	1.68	-4.23	29	C 29	-0.96	1.43	-1.19	63	C 63	-0.48	52	1.94	-1.41	0.90	0.07	1.39	0.44
34	C 34	0.54	0.14	1.86	-5.37	17	C 17	-0.87	1.27	-0.84	124	C124	1.51	53	0.44	-1.29	0.92	0.06	1.23	0.44
35	C 35	-1.04	0.17	1.67	-1.59	47	C 47	-0.87	1.12	-0.18	29	C 29	-0.96	33	1.71	-1.19	0.89	0.09	1.43	0.43
36	C 36	-1.07	0.17	1.68	-1.65	130	C130	-0.85	1.26	-0.44	72	C 72	0.24	33	-0.57	-1.15	0.94	0.05	1.14	0.44
37	C 37	-3.02	0.36	1.59	-0.04	5	C 5	-0.79	1.64	-1.81	99	C 99	0.32	7	1.16	-1.07	0.95	0.05	1.18	0.43
38	C 38	-0.03	0.14	0.58	1.76	112	C112	-0.77	1.14	-0.43	42	C 42	-0.46	49	-0.34	-1.01	0.93	0.07	1.21	0.41
39	C 39	1.61	0.15	0.26	2.69	98	C 98	-0.74	0.78	0.52	122	C122	1.72	46	3.63	-0.98	0.93	0.07	1.06	0.38
40	C 40	1.30	0.14	1.36	-2.10	78	C 78	-0.74	0.66	0.84	128	C128	-0.17	49	1.65	-0.91	0.95	0.06	1.27	0.43
41	C 41	0.62	0.14	0.99	0.00	43	C 43	-0.74	0.67	0.80	101	C101	-0.67	53	0.17	-0.91	0.93	0.08	1.28	0.40
42	C 42	-0.46	0.15	1.21	-1.01	28	C 28	-0.72	1.65	-1.86	7	C 7	-1.40	44	1.27	-0.88	0.89	0.12	1.51	0.40
43	C 43	-0.74	0.16	0.67	0.80	126	C126	-0.69	1.04	0.14	2	C 2	-1.01	39	0.08	-0.85	0.92	0.10	1.33	0.39
44	C 44	1.26	0.14	0.76	1.10	101	C101	-0.67	1.28	-0.91	52	C 52	-1.77	50	3.26	-0.85	0.86	0.16	1.80	0.41
45	C 45	0.45	0.14	0.88	1.06	71	C 71	-0.64	0.94	0.24	17	C 17	-0.87	53	-0.38	-0.84	0.93	0.09	1.27	0.39
46	C 46	1.49	0.14	0.58	1.65	26	C 26	-0.62	0.88	0.34	60	C 60	-0.28	47	0.88	-0.83	0.95	0.06	1.23	0.41
47	C 47	-0.87	0.17	1.12	-0.18	63	C 63	-0.48	1.39	-1.41	54	C 54	-1.10	36	0.89	-0.79	0.92	0.10	1.40	0.38
48	C 48	-1.44	0.19	1.43	-0.73	42	C 42	-0.46	1.21	-1.01	48	C 48	-1.44	26	0.92	-0.73	0.91	0.13	1.43	0.38
49	C 49	1.07	0.14	1.00	-0.51	32	C 32	-0.41	1.37	-1.45	14	C 14	0.96	51	-1.04	-0.70	0.96	0.05	1.10	0.42
50	C 50	-0.09	0.14	1.67	-3.32	4	C 4	-0.39	1.61	-2.65	64	C 64	-1.97	48	2.52	-0.67	0.88	0.18	1.72	0.39

TABLE CONTINUED

SEQ NUM	ITEM NAME	ITEM DIFF	STD ERROR	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	INFO AMT	FIT T-TESTS BETWN	FIT T-TESTS TOTAL	WTD MNSQ	MNSQ SD	DISC INDX	POINT BISER
51	C 51	1.07	0.14	0.92	0.34	56	C 56	-0.35	1.78	-3.47	68	C 68	-1.44	51	1.14	-0.65	0.91	0.13	1.41	0.37
52	C 52	-1.77	0.22	1.80	-0.85	140	C140	-0.30	1.09	-0.45	53	C 53	0.13	21	-1.11	-0.63	0.97	0.05	1.13	0.40
53	C 53	0.13	0.14	1.13	-0.63	133	C133	-0.28	0.54	1.60	18	C 18	-1.77	51	1.68	-0.63	0.90	0.16	1.56	0.36
54	C 54	-1.10	0.18	1.40	-0.79	60	C 60	-0.28	1.23	-0.83	80	C 80	-0.03	32	-0.58	-0.62	0.97	0.05	1.02	0.39
55	C 55	0.58	0.14	0.46	3.10	61	C 61	-0.22	1.33	-1.60	109	C109	-0.07	53	0.45	-0.54	0.97	0.06	1.16	0.40
56	C 56	-0.35	0.15	1.78	-3.47	79	C 79	-0.20	0.68	1.28	102	C102	-1.64	45	1.45	-0.52	0.92	0.14	1.45	0.36
57	C 57	1.30	0.14	1.51	-2.89	121	C121	-0.20	0.89	0.45	49	C 49	1.07	49	1.55	-0.51	0.97	0.05	1.00	0.39
58	C 58	0.20	0.14	0.94	0.39	128	C128	-0.17	1.27	-0.91	113	C113	0.88	51	1.55	-0.47	0.98	0.05	1.08	0.42
59	C 59	0.86	0.14	0.57	2.65	65	C 65	-0.17	1.37	-1.79	140	C140	-0.30	52	0.08	-0.45	0.97	0.06	1.09	0.38
60	C 60	-0.28	0.15	1.23	-0.83	123	C123	-0.11	0.40	2.85	130	C130	-0.85	46	0.62	-0.44	0.96	0.09	1.26	0.36
61	C 61	-0.22	0.15	1.33	-1.60	50	C 50	-0.09	1.67	-3.32	112	C112	-0.77	47	-0.76	-0.43	0.96	0.08	1.14	0.34
62	C 62	0.41	0.14	1.85	-5.41	109	C109	-0.07	1.16	-0.54	20	C 20	1.79	52	0.15	-0.43	0.97	0.07	0.99	0.36
63	C 63	-0.48	0.15	1.39	-1.41	80	C 80	-0.03	1.02	-0.62	23	C 23	-1.51	43	0.32	-0.39	0.94	0.13	1.32	0.32
64	C 64	-1.97	0.23	1.72	-0.67	38	C 38	-0.03	0.58	1.76	100	C100	-1.26	18	-0.26	-0.37	0.95	0.11	1.05	0.30
65	C 65	-0.17	0.14	1.37	-1.79	16	C 16	0.01	1.59	-3.00	31	C 31	1.96	47	1.27	-0.34	0.97	0.08	1.04	0.36
66	C 66	-1.82	0.22	1.18	-0.20	81	C 81	0.05	0.91	0.06	104	C104	-2.14	20	1.95	-0.32	0.93	0.20	1.46	0.31
67	C 67	-1.17	0.18	1.74	-1.59	83	C 83	0.05	0.64	1.80	138	C138	0.30	31	-0.18	-0.22	0.99	0.05	1.02	0.39
68	C 68	-1.44	0.19	1.41	-0.65	53	C 53	0.13	1.13	-0.63	66	C 66	-1.82	26	0.08	-0.20	0.96	0.16	1.18	0.26
69	C 69	4.59	0.39	-0.86	0.35	58	C 58	0.20	0.94	0.39	47	C 47	-0.87	6	-1.13	-0.18	0.98	0.09	1.12	0.33
70	C 70	-2.28	0.26	1.13	-0.14	127	C127	0.20	1.63	-3.66	11	C 11	-1.97	14	-0.50	-0.18	0.96	0.18	1.20	0.25
71	C 71	-0.64	0.16	0.94	0.24	6	C 6	0.20	1.61	-3.69	70	C 70	-2.28	40	-0.35	-0.14	0.96	0.21	1.13	0.23
72	C 72	0.24	0.14	1.14	-1.15	84	C 84	0.22	0.79	1.23	37	C 37	-3.02	52	0.25	-0.04	0.95	0.32	1.59	0.21
73	C 73	1.21	0.14	0.75	1.31	33	C 33	0.24	1.68	-4.23	116	C116	1.05	50	2.07	-0.04	1.00	0.05	0.96	0.37
74	C 74	1.13	0.14	-0.05	5.42	139	C139	0.24	0.84	0.77	24	C 24	-1.72	51	-0.66	-0.02	0.99	0.15	1.03	0.23
75	C 75	0.30	0.14	0.34	3.43	72	C 72	0.24	1.14	-1.15	41	C 41	0.62	52	0.16	0.00	1.00	0.05	0.99	0.38
76	C 76	0.41	0.14	0.21	4.36	19	C 19	0.26	1.69	-3.80	93	C 93	-3.73	52	-0.44	0.06	0.95	0.47	1.49	0.16
77	C 77	2.81	0.19	-0.10	1.62	75	C 75	0.30	0.34	3.43	90	C 90	-3.32	26	0.40	0.06	0.98	0.38	0.84	0.09
78	C 78	-0.74	0.16	0.66	0.84	138	C138	0.30	1.02	-0.22	81	C 81	0.05	39	-0.32	0.06	1.00	0.05	0.91	0.36
79	C 79	-0.20	0.15	0.68	1.28	99	C 99	0.32	1.18	-1.07	91	C 91	-2.89	47	0.00	0.09	1.00	0.30	0.69	0.08
80	C 80	-0.03	0.14	1.02	-0.62	114	C114	0.34	1.27	-1.84	82	C 82	-1.92	49	-1.48	0.11	1.01	0.17	0.93	0.17
81	C 81	0.05	0.14	0.91	0.06	1	C 1	0.37	0.62	2.07	89	C 89	-4.02	50	-0.62	0.11	0.96	0.55	0.83	0.08
82	C 82	-1.92	0.23	0.93	0.11	62	C 62	0.41	1.85	-5.41	92	C 92	-3.32	19	-0.29	0.12	1.00	0.38	0.73	0.05
83	C 83	0.05	0.14	0.64	1.80	76	C 76	0.41	0.21	4.36	21	C 21	-2.08	50	0.24	0.13	1.01	0.19	0.89	0.16
84	C 84	0.22	0.14	0.79	1.23	45	C 45	0.45	0.88	1.06	126	C126	-0.69	51	0.54	0.14	1.01	0.08	1.04	0.32
85	C 85	-1.55	0.20	0.51	0.54	13	C 13	0.47	0.93	0.44	88	C 88	-2.68	24	0.58	0.15	1.02	0.27	0.76	0.09
86	C 86	-1.30	0.19	0.38	1.03	95	C 95	0.54	0.97	0.43	71	C 71	-0.64	29	-0.70	0.24	1.02	0.08	0.94	0.29
87	C 87	-1.10	0.18	0.09	1.67	34	C 34	0.54	1.86	-5.37	94	C 94	-5.13	32	-1.37	0.28	0.96	0.97	1.71	0.10
88	C 88	-2.68	0.31	0.76	0.15	55	C 55	0.58	0.46	3.10	26	C 26	-0.62	10	-0.42	0.34	1.02	0.09	0.88	0.29
89	C 89	-4.02	0.58	0.83	0.11	125	C125	0.58	0.90	0.79	51	C 51	1.07	3	1.79	0.34	1.02	0.05	0.92	0.35
90	C 90	-3.32	0.42	0.84	0.06	3	C 3	0.60	1.34	-2.28	69	C 69	4.59	5	9.84	0.35	1.08	0.35	-0.86	-0.22
91	C 91	-2.89	0.34	0.69	0.09	41	C 41	0.62	0.99	0.00	58	C 58	0.20	8	0.74	0.39	1.02	0.05	0.94	0.35
92	C 92	-3.32	0.42	0.73	0.12	96	C 96	0.79	1.59	-3.47	95	C 95	0.54	5	-0.67	0.43	1.02	0.05	0.97	0.36
93	C 93	-3.73	0.51	1.49	0.06	107	C107	0.79	1.51	-3.46	13	C 13	0.47	4	-1.37	0.44	1.02	0.05	0.93	0.36
94	C 94	-5.13	1.00	1.71	0.28	97	C 97	0.81	-0.09	5.69	121	C121	-0.20	1	-0.05	0.45	1.03	0.06	0.89	0.31
95	C 95	0.54	0.14	0.97	0.43	10	C 10	0.84	1.37	-2.63	98	C 98	-0.74	53	-0.47	0.52	1.04	0.08	0.78	0.25
96	C 96	0.79	0.14	1.59	-3.47	103	C103	0.84	1.97	-5.90	85	C 85	-1.55	53	2.13	0.54	1.07	0.14	0.51	0.12
97	C 97	0.81	0.14	-0.09	5.69	59	C 59	0.86	0.57	2.65	25	C 25	1.30	53	-0.19	0.65	1.04	0.06	0.78	0.32
98	C 98	-0.74	0.16	0.78	0.52	113	C113	0.88	1.08	-0.47	27	C 27	3.05	39	3.63	0.65	1.09	0.15	0.39	0.08
99	C 99	0.32	0.14	1.18	-1.07	30	C 30	0.90	-0.24	6.13	8	C 8	1.23	52	0.38	0.73	1.04	0.06	0.94	0.34
100	C100	-1.26	0.18	1.05	-0.37	118	C118	0.92	-0.08	5.76	117	C117	2.96	30	2.40	0.77	1.10	0.14	0.54	0.11
101	C101	-0.67	0.16	1.28	-0.91	131	C131	0.94	0.25	4.27	139	C139	0.24	40	2.24	0.77	1.04	0.05	0.84	0.33

****

TABLE CONTINUED

SEQ NUM	ITEM NAME	ITEM DIFF	STD ERROR	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	INFO AMT	FIT T-TESTS BETWN	FIT T-TESTS TOTAL	WTD MNSQ MNSQ	WTD MNSQ SD	DISC POINT INDX	DISC POINT BISER
102	C102	-1.64	0.21	1.45	-0.52	14	C 14	0.96	1.10	-0.70	135	C135	2.57	23	1.25	0.77	1.08	0.11	0.61	0.17
103	C103	0.84	0.14	1.97	-5.90	105	C105	1.01	1.26	-1.72	125	C125	0.58	53	0.10	0.79	1.04	0.05	0.90	0.34
104	C104	-2.14	0.25	1.46	-0.32	116	C116	1.05	0.96	-0.04	43	C 43	-0.74	16	0.94	0.80	1.06	0.08	0.67	0.22
105	C105	1.01	0.14	1.26	-1.72	51	C 51	1.07	0.92	0.34	78	C 78	-0.74	52	0.77	0.84	1.07	0.08	0.66	0.22
106	C106	1.13	0.14	1.57	-2.95	49	C 49	1.07	1.00	-0.51	137	C137	2.32	51	1.02	0.99	1.09	0.10	0.58	0.19
107	C107	0.79	0.14	1.51	-3.46	15	C 15	1.07	1.66	-4.10	132	C132	1.68	53	1.89	1.03	1.07	0.07	0.67	0.25
108	C108	1.26	0.14	0.62	2.15	9	C 9	1.09	0.42	3.08	86	C 86	-1.30	50	2.39	1.03	1.12	0.12	0.38	0.08
109	C109	-0.07	0.14	1.16	-0.54	12	C 12	1.09	1.60	-3.69	45	C 45	0.45	49	0.73	1.06	1.05	0.05	0.88	0.32
110	C110	1.49	0.14	0.46	2.35	106	C106	1.13	1.57	-2.95	134	C134	1.32	47	0.75	1.07	1.06	0.06	0.74	0.30
111	C111	1.79	0.15	0.23	2.84	74	C 74	1.13	-0.05	5.42	44	C 44	1.26	43	0.26	1.10	1.06	0.06	0.76	0.30
112	C112	-0.77	0.16	1.14	-0.43	73	C 73	1.21	0.75	1.31	84	C 84	0.22	38	0.16	1.23	1.06	0.05	0.79	0.30
113	C113	0.88	0.14	1.08	-0.47	8	C 8	1.23	0.94	0.73	115	C115	1.66	52	3.37	1.24	1.08	0.07	0.56	0.23
114	C114	0.34	0.14	1.27	-1.84	44	C 44	1.26	0.76	1.10	79	C 79	-0.20	52	1.19	1.28	1.08	0.06	0.68	0.26
115	C115	1.66	0.15	0.56	1.24	108	C108	1.26	0.62	2.15	73	C 73	1.21	45	1.59	1.31	1.07	0.05	0.75	0.28
116	C116	1.05	0.14	0.96	-0.04	120	C120	1.26	0.44	3.01	133	C133	-0.28	52	1.91	1.60	1.10	0.06	0.54	0.21
117	C117	2.96	0.20	0.54	0.77	136	C136	1.28	0.65	2.18	77	C 77	2.81	24	5.31	1.62	1.21	0.13	-0.10	-0.03
118	C118	0.92	0.14	-0.08	5.76	40	C 40	1.30	1.36	-2.10	46	C 46	1.49	52	2.22	1.65	1.10	0.06	0.58	0.23
119	C119	1.34	0.14	-0.13	4.72	25	C 25	1.30	0.78	0.65	87	C 87	-1.10	49	3.87	1.67	1.18	0.10	0.09	0.01
120	C120	1.26	0.14	0.44	3.01	57	C 57	1.30	1.51	-2.89	38	C 38	-0.03	50	1.61	1.76	1.10	0.05	0.58	0.25
121	C121	-0.20	0.15	0.89	0.45	134	C134	1.32	0.74	1.07	83	C 83	0.05	47	0.44	1.80	1.10	0.05	0.64	0.25
122	C122	1.72	0.15	1.06	-0.98	119	C119	1.34	-0.13	4.72	1	C 1	0.37	44	1.76	2.07	1.10	0.05	0.62	0.26
123	C123	-0.11	0.14	0.40	2.85	110	C110	1.49	0.46	2.35	108	C108	1.26	48	3.04	2.15	1.12	0.06	0.62	0.27
124	C124	1.51	0.15	1.23	-1.29	46	C 46	1.49	0.58	1.65	136	C136	1.28	47	1.34	2.18	1.13	0.06	0.65	0.25
125	C125	0.58	0.14	0.90	0.79	124	C124	1.51	1.23	-1.29	110	C110	1.49	53	2.13	2.35	1.15	0.06	0.46	0.22
126	C126	-0.69	0.16	1.04	0.14	22	C 22	1.55	1.48	-2.43	59	C 59	0.86	40	1.85	2.65	1.14	0.05	0.57	0.23
127	C127	0.20	0.14	1.63	-3.66	39	C 39	1.61	0.26	2.69	39	C 39	1.61	51	3.43	2.69	1.18	0.07	0.26	0.15
128	C128	-0.17	0.14	1.27	-0.91	115	C115	1.66	0.56	1.24	111	C111	1.79	47	3.98	2.84	1.21	0.07	0.23	0.09
129	C129	1.77	0.15	1.58	-2.70	132	C132	1.68	0.67	1.03	123	C123	-0.11	44	2.63	2.85	1.17	0.06	0.40	0.16
130	C130	-0.85	0.16	1.26	-0.44	122	C122	1.72	1.06	-0.98	120	C120	1.26	37	1.96	3.01	1.18	0.06	0.44	0.19
131	C131	0.94	0.14	0.25	4.27	129	C129	1.77	1.58	-2.70	9	C 9	1.09	52	2.78	3.08	1.17	0.05	0.42	0.21
132	C132	1.68	0.15	0.67	1.03	20	C 20	1.79	0.99	-0.43	55	C 55	0.58	45	4.63	3.10	1.16	0.05	0.46	0.21
133	C133	-0.28	0.15	0.54	1.60	111	C111	1.79	0.23	2.84	75	C 75	0.30	46	2.30	3.43	1.18	0.05	0.34	0.17
134	C134	1.32	0.14	0.74	1.07	31	C 31	1.96	1.04	-0.34	131	C131	0.94	49	4.13	4.27	1.23	0.05	0.25	0.15
135	C135	2.57	0.18	0.61	0.77	137	C137	2.32	0.58	0.99	76	C 76	0.41	30	3.75	4.36	1.23	0.05	0.21	0.13
136	C136	1.28	0.14	0.65	2.18	135	C135	2.57	0.61	0.77	119	C119	1.34	50	6.30	4.72	1.30	0.06	-0.13	0.03
137	C137	2.32	0.17	0.58	0.99	77	C 77	2.81	-0.10	1.62	74	C 74	1.13	35	4.96	5.42	1.32	0.05	-0.05	0.05
138	C138	0.30	0.14	1.02	-0.22	117	C117	2.96	0.54	0.77	97	C 97	0.81	52	5.77	5.69	1.31	0.05	-0.09	0.05
139	C139	0.24	0.14	0.84	0.77	27	C 27	3.05	0.39	0.65	118	C118	0.92	52	5.03	5.76	1.32	0.05	-0.08	0.03
140	C140	-0.30	0.15	1.09	-0.45	69	C 69	4.59	-0.86	0.35	30	C 30	0.90	46	6.19	6.13	1.34	0.05	-0.24	0.01
	MEAN	0.00		1.02	-0.13										1.57	-0.13	0.99	0.10		
	S.D.	1.47		0.52	2.18										1.80	2.18	0.13	0.11		

140 ITEMS CALIBRATED ON 251 PERSONS
251 MEASURABLE PERSONS WITH MEAN ABILITY = 0.69 AND STD. DEV. = 0.85

SEQ NUM	ITEM NAME	ITEM DIFF	STD ERROR	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	INFO AMT	FIT BETWN	T-TESTS TOTAL	WTD MNSQ	MNSQ SD	DISC INDX	POINT BISER
1	D 1	-1.90	0.21	1.39	-0.26	98	D 98	-3.53	1.09	0.02	36	D 36	0.35	22	4.36	-5.90	0.76	0.04	2.03	0.65
2	D 2	-1.29	0.17	1.70	-1.13	22	D 22	-2.64	1.35	-0.13	66	D 66	0.06	33	4.55	-5.82	0.76	0.04	2.11	0.64
3	D 3	1.40	0.15	1.04	-0.31	9	D 9	-2.42	1.58	-0.18	45	D 45	-0.07	42	3.22	-4.35	0.81	0.05	1.87	0.57
4	D 4	-0.39	0.14	1.55	-2.58	19	D 19	-2.30	1.57	-0.29	32	D 32	0.01	47	2.81	-3.93	0.83	0.04	1.76	0.55
5	D 5	0.06	0.14	1.55	-2.57	48	D 48	-1.99	2.00	-0.80	111	D111	0.39	51	2.37	-3.80	0.84	0.04	1.73	0.54
6	D 6	1.89	0.17	0.68	0.46	33	D 33	-1.99	1.96	-0.90	4	D 4	-0.39	34	1.24	-2.58	0.87	0.05	1.55	0.49
7	D 7	-1.73	0.20	1.88	-1.04	1	D 1	-1.90	1.39	-0.26	5	D 5	0.06	25	2.80	-2.57	0.89	0.04	1.55	0.48
8	D 8	-1.45	0.18	1.55	-0.82	63	D 63	-1.77	1.88	-0.79	80	D 80	-0.44	30	2.67	-2.53	0.87	0.06	1.67	0.49
9	D 9	-2.42	0.25	1.58	-0.18	51	D 51	-1.77	1.26	-0.34	110	D110	1.13	15	3.13	-2.43	0.86	0.06	1.62	0.49
10	D 10	-0.86	0.16	0.17	1.81	7	D 7	-1.73	1.88	-1.04	78	D 78	-0.61	40	2.83	-2.39	0.86	0.06	1.67	0.49
11	D 11	1.40	0.15	0.74	-0.16	34	D 34	-1.73	1.43	-0.47	26	D 26	0.03	42	1.75	-2.31	0.90	0.04	1.50	0.47
12	D 12	0.28	0.14	1.06	-0.47	119	D119	-1.73	0.79	0.16	72	D 72	-0.99	52	3.37	-2.30	0.82	0.08	1.91	0.53
13	D 13	-0.89	0.16	1.28	-0.83	18	D 18	-1.66	1.53	-0.39	21	D 21	0.76	40	2.89	-2.30	0.89	0.05	1.44	0.47
14	D 14	1.65	0.16	0.40	1.50	96	D 96	-1.58	1.72	-1.03	27	D 27	-0.70	38	2.50	-2.26	0.86	0.07	1.71	0.50
15	D 15	-1.55	0.19	1.58	-0.89	16	D 16	-1.55	1.48	-0.52	56	D 56	1.02	28	1.82	-2.17	0.88	0.06	1.46	0.47
16	D 16	-1.55	0.19	1.48	-0.52	15	D 15	-1.55	1.58	-0.89	68	D 68	-1.38	28	4.43	-2.13	0.79	0.11	2.11	0.55
17	D 17	-1.18	0.17	1.52	-1.17	100	D100	-1.51	1.41	-0.46	128	D128	-0.37	35	1.86	-2.11	0.89	0.05	1.56	0.47
18	D 18	-1.66	0.19	1.53	-0.39	8	D 8	-1.45	1.55	-0.82	40	D 40	-1.04	26	2.61	-1.93	0.84	0.08	1.74	0.49
19	D 19	-2.30	0.24	1.57	-0.29	68	D 68	-1.38	2.11	-2.13	135	D135	1.00	17	2.12	-1.88	0.90	0.06	1.45	0.45
20	D 20	-1.26	0.17	0.82	0.23	31	D 31	-1.32	1.77	-1.38	109	D109	0.49	33	1.20	-1.76	0.92	0.04	1.37	0.44
21	D 21	0.76	0.14	1.44	-2.30	25	D 25	-1.29	0.54	0.96	92	D 92	-0.44	50	0.45	-1.76	0.90	0.06	1.44	0.44
22	D 22	-2.64	0.28	1.35	-0.13	2	D 2	-1.29	1.70	-1.13	108	D108	0.06	13	1.26	-1.75	0.92	0.04	1.35	0.43
23	D 23	-0.33	0.14	0.42	2.32	20	D 20	-1.26	0.82	0.23	52	D 52	-0.91	48	2.22	-1.64	0.88	0.08	1.76	0.46
24	D 24	-0.35	0.14	1.46	-1.56	28	D 28	-1.23	1.44	-0.72	24	D 24	-0.35	48	1.85	-1.56	0.92	0.05	1.46	0.44
25	D 25	-1.29	0.17	0.54	0.96	82	D 82	-1.23	1.58	-1.22	117	D117	1.00	33	0.97	-1.44	0.92	0.06	1.33	0.43
26	D 26	0.03	0.14	1.50	-2.31	17	D 17	-1.18	1.52	-1.17	65	D 65	-1.04	51	1.86	-1.40	0.89	0.08	1.72	0.45
27	D 27	-0.70	0.15	1.71	-2.26	81	D 81	-1.15	1.07	-0.20	31	D 31	-1.32	43	2.02	-1.38	0.86	0.10	1.77	0.46
28	D 28	-1.23	0.17	1.44	-0.72	132	D132	-1.12	1.57	-1.37	132	D132	-1.12	34	1.49	-1.37	0.88	0.09	1.57	0.43
29	D 29	-0.79	0.15	1.18	-0.24	59	D 59	-1.07	1.34	-0.69	39	D 39	-0.63	42	1.64	-1.23	0.92	0.06	1.34	0.41
30	D 30	-0.15	0.14	1.30	-1.12	43	D 43	-1.07	0.83	0.04	82	D 82	-1.23	50	2.97	-1.22	0.88	0.10	1.58	0.40
31	D 31	-1.32	0.17	1.77	-1.38	40	D 40	-1.04	1.74	-1.93	17	D 17	-1.18	32	0.97	-1.17	0.89	0.09	1.52	0.41
32	D 32	0.01	0.14	1.76	-3.93	65	D 65	-1.04	1.72	-1.40	138	D138	1.47	51	2.44	-1.16	0.91	0.08	1.22	0.38
33	D 33	-1.99	0.22	1.96	-0.90	70	D 70	-1.01	1.43	-0.73	2	D 2	-1.29	21	2.06	-1.13	0.89	0.10	1.70	0.43
34	D 34	-1.73	0.20	1.43	-0.47	72	D 72	-0.99	1.91	-2.30	41	D 41	0.22	25	1.06	-1.13	0.95	0.04	1.19	0.41
35	D 35	0.70	0.14	1.03	-0.35	61	D 61	-0.94	-0.09	2.09	30	D 30	-0.15	50	1.32	-1.12	0.95	0.05	1.30	0.40
36	D 36	0.35	0.14	2.03	-5.90	52	D 52	-0.91	1.76	-1.64	69	D 69	0.64	52	0.46	-1.11	0.95	0.05	1.30	0.41
37	D 37	-0.46	0.15	1.01	0.05	13	D 13	-0.89	1.28	-0.83	122	D122	0.20	47	1.11	-1.11	0.95	0.04	1.26	0.40
38	D 38	2.79	0.22	0.57	0.07	10	D 10	-0.86	0.17	1.81	79	D 79	-0.77	19	0.75	-1.07	0.93	0.07	1.46	0.40
39	D 39	-0.63	0.15	1.34	-1.23	29	D 29	-0.79	1.18	-0.24	7	D 7	-1.73	44	2.49	-1.04	0.86	0.13	1.88	0.44
40	D 40	-1.04	0.16	1.74	-1.93	120	D120	-0.77	1.30	-0.82	96	D 96	-1.58	37	1.76	-1.03	0.88	0.12	1.72	0.42
41	D 41	0.22	0.14	1.19	-1.13	79	D 79	-0.77	1.46	-1.07	121	D121	0.18	51	0.80	-1.01	0.96	0.04	1.12	0.40
42	D 42	-0.31	0.14	0.90	0.53	47	D 47	-0.74	1.16	-0.80	62	D 62	1.28	48	3.13	-0.94	0.94	0.07	1.22	0.36
43	D 43	-1.07	0.16	0.83	0.04	27	D 27	-0.70	1.71	-2.26	33	D 33	-1.99	37	2.83	-0.90	0.86	0.16	1.96	0.43
44	D 44	1.47	0.16	1.06	0.24	49	D 49	-0.70	1.05	0.26	15	D 15	-1.55	41	0.93	-0.89	0.89	0.12	1.58	0.40
45	D 45	-0.07	0.14	1.87	-4.35	39	D 39	-0.63	1.34	-1.23	86	D 86	-0.23	50	0.49	-0.84	0.96	0.05	1.14	0.38
46	D 46	0.22	0.14	0.34	3.42	84	D 84	-0.63	-0.05	2.78	13	D 13	-0.89	51	0.14	-0.83	0.94	0.08	1.28	0.37
47	D 47	-0.74	0.15	1.16	-0.80	78	D 78	-0.61	1.67	-2.39	8	D 8	-1.45	42	3.01	-0.82	0.91	0.11	1.55	0.39
48	D 48	-1.99	0.22	2.00	-0.80	75	D 75	-0.48	0.85	0.15	120	D120	-0.77	21	-0.48	-0.82	0.94	0.07	1.30	0.37
49	D 49	-0.70	0.15	1.05	0.26	37	D 37	-0.46	1.01	0.05	47	D 47	-0.74	43	0.58	-0.80	0.95	0.07	1.16	0.36
50	D 50	0.98	0.14	0.59	1.54	80	D 80	-0.44	1.67	-2.53	48	D 48	-1.99	47	3.05	-0.80	0.87	0.16	2.00	0.41

****

TABLE CONTINUED

SEQ NUM	ITEM NAME	ITEM DIFF	STD ERROR	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	INFO AMT	FIT BETWN	T-TESTS TOTAL	WTD MNSQ	MNSQ SD	DISC INDX	POINT BISER
51	D 51	-1.77	0.20	1.26	-0.34	92	D 92	-0.44	1.44	-1.76	63	D 63	-1.77	25	2.61	-0.79	0.89	0.14	1.88	0.40
52	D 52	-0.91	0.16	1.76	-1.64	91	D 91	-0.41	1.11	-0.28	113	D113	1.45	40	0.38	-0.77	0.94	0.07	1.24	0.38
53	D 53	1.11	0.15	0.56	1.35	106	D106	-0.39	0.89	0.48	70	D 70	-1.01	46	0.48	-0.73	0.94	0.08	1.43	0.38
54	D 54	-0.13	0.14	0.55	2.33	4	D 4	-0.39	1.55	-2.58	112	D112	-0.27	50	0.51	-0.72	0.96	0.05	1.16	0.37
55	D 55	2.48	0.20	0.52	0.45	128	D128	-0.37	1.56	-2.11	28	D 28	-1.23	24	0.42	-0.72	0.93	0.10	1.44	0.37
56	D 56	1.02	0.15	1.46	-2.17	71	D 71	-0.37	1.18	-0.57	59	D 59	-1.07	47	0.77	-0.69	0.94	0.09	1.34	0.36
57	D 57	0.39	0.14	0.47	2.41	24	D 24	-0.35	1.46	-1.56	115	D115	0.41	51	1.27	-0.68	0.97	0.04	1.11	0.38
58	D 58	0.56	0.14	1.12	-0.65	95	D 95	-0.35	0.65	1.42	58	D 58	0.56	51	1.86	-0.65	0.97	0.05	1.12	0.37
59	D 59	-1.07	0.16	1.34	-0.69	23	D 23	-0.33	0.42	2.32	71	D 71	-0.37	37	-0.40	-0.57	0.97	0.05	1.18	0.36
60	D 60	0.92	0.14	1.02	-0.17	131	D131	-0.33	0.34	2.93	16	D 16	-1.55	48	0.57	-0.52	0.94	0.12	1.48	0.33
61	D 61	-0.94	0.16	-0.09	2.09	83	D 83	-0.33	0.54	1.58	12	D 12	0.28	39	1.20	-0.47	0.98	0.04	1.06	0.37
62	D 62	1.28	0.15	1.22	-0.94	42	D 42	-0.31	0.90	0.53	34	D 34	-1.73	44	0.36	-0.47	0.93	0.13	1.43	0.33
63	D 63	-1.77	0.20	1.88	-0.79	112	D112	-0.27	1.16	-0.72	100	D100	-1.51	25	0.52	-0.46	0.94	0.12	1.41	0.34
64	D 64	2.85	0.23	0.55	0.04	101	D101	-0.27	0.76	1.09	18	D 18	-1.66	18	1.40	-0.39	0.95	0.13	1.53	0.34
65	D 65	-1.04	0.16	1.72	-1.40	86	D 86	-0.23	1.14	-0.84	125	D125	1.78	37	-0.22	-0.38	0.96	0.09	1.08	0.32
66	D 66	0.06	0.14	2.11	-5.82	118	D118	-0.21	0.82	0.65	35	D 35	0.70	51	0.09	-0.35	0.98	0.05	1.03	0.37
67	D 67	0.74	0.14	0.26	2.94	88	D 88	-0.17	0.18	3.55	51	D 51	-1.77	50	0.11	-0.34	0.95	0.14	1.26	0.29
68	D 68	-1.38	0.18	2.11	-2.13	87	D 87	-0.17	0.08	3.77	3	D 3	1.40	31	0.48	-0.31	0.98	0.07	1.04	0.35
69	D 69	0.64	0.14	1.30	-1.11	30	D 30	-0.15	1.30	-1.12	19	D 19	-2.30	50	1.85	-0.29	0.93	0.19	1.57	0.28
70	D 70	-1.01	0.16	1.43	-0.73	54	D 54	-0.13	0.55	2.33	91	D 91	-0.41	38	-0.82	-0.28	0.98	0.05	1.11	0.34
71	D 71	-0.37	0.14	1.18	-0.57	127	D127	-0.09	0.85	0.97	1	D 1	-1.90	48	-0.37	-0.26	0.96	0.15	1.39	0.28
72	D 72	-0.99	0.16	1.91	-2.30	45	D 45	-0.07	1.87	-4.35	29	D 29	-0.79	38	0.78	-0.24	0.98	0.07	1.18	0.33
73	D 73	3.29	0.27	-0.64	0.74	90	D 90	-0.07	1.02	0.10	81	D 81	-1.15	13	0.14	-0.20	0.98	0.09	1.07	0.29
74	D 74	-0.05	0.14	0.71	0.92	74	D 74	-0.05	0.71	0.92	9	D 9	-2.42	50	1.28	-0.18	0.95	0.20	1.58	0.27
75	D 75	-0.46	0.15	0.85	0.15	32	D 32	0.01	1.76	-3.93	60	D 60	0.92	46	0.32	-0.17	0.99	0.05	1.02	0.34
76	D 76	1.78	0.17	0.41	1.53	26	D 26	0.03	1.50	-2.31	11	D 11	1.40	36	5.78	-0.16	0.99	0.07	0.74	0.25
77	D 77	0.70	0.14	0.22	3.87	108	D108	0.06	1.35	-1.75	102	D102	0.74	50	-1.70	-0.15	0.99	0.05	0.95	0.35
78	D 78	-0.61	0.15	1.67	-2.39	66	D 66	0.06	2.11	-5.82	126	D126	0.47	44	-1.29	-0.14	0.99	0.04	0.94	0.35
79	D 79	-0.77	0.15	1.46	-1.07	5	D 5	0.06	1.55	-2.57	22	D 22	-2.64	42	-0.48	-0.13	0.95	0.23	1.35	0.22
80	D 80	-0.44	0.15	1.67	-2.53	121	D121	0.18	1.12	-1.01	123	D123	0.68	47	0.41	-0.08	1.00	0.05	1.09	0.34
81	D 81	-1.15	0.17	1.07	-0.20	122	D122	0.20	1.26	-1.11	98	D 98	-3.53	36	0.07	0.02	0.96	0.38	1.09	0.13
82	D 82	-1.23	0.17	1.58	-1.22	46	D 46	0.22	0.34	3.42	64	D 64	2.85	34	2.74	0.04	1.00	0.17	0.55	0.11
83	D 83	-0.33	0.14	0.54	1.58	41	D 41	0.22	1.19	-1.13	43	D 43	-1.07	48	0.29	0.04	1.00	0.09	0.83	0.25
84	D 84	-0.63	0.15	-0.05	2.78	114	D114	0.24	0.55	2.21	37	D 37	-0.46	44	-0.02	0.05	1.00	0.06	1.01	0.32
85	D 85	0.49	0.14	-1.03	8.91	12	D 12	0.28	1.06	-0.47	38	D 38	2.79	51	3.11	0.07	1.00	0.17	0.57	0.12
86	D 86	-0.23	0.14	1.14	-0.84	36	D 36	0.35	2.03	-5.90	90	D 90	-0.07	49	-0.62	0.10	1.00	0.05	1.02	0.33
87	D 87	-0.17	0.14	0.08	3.77	111	D111	0.39	1.73	-3.80	75	D 75	-0.48	50	0.49	0.15	1.01	0.06	0.85	0.30
88	D 88	-0.17	0.14	0.18	3.55	57	D 57	0.39	0.47	2.41	119	D119	-1.73	50	0.03	0.16	1.02	0.13	0.79	0.17
89	D 89	0.90	0.14	0.93	0.50	115	D115	0.41	1.11	-0.68	136	D136	2.08	48	-0.79	0.18	1.02	0.11	0.89	0.25
90	D 90	-0.07	0.14	1.02	0.10	126	D126	0.47	0.94	-0.14	20	D 20	-1.26	50	0.13	0.23	1.02	0.10	0.82	0.21
91	D 91	-0.41	0.15	1.11	-0.28	85	D 85	0.49	-1.03	8.91	44	D 44	1.47	47	-0.83	0.24	1.02	0.08	1.06	0.32
92	D 92	-0.44	0.15	1.44	-1.76	109	D109	0.49	1.37	-1.76	49	D 49	-0.70	47	1.23	0.26	1.02	0.07	1.05	0.29
93	D 93	1.26	0.15	0.84	0.42	58	D 58	0.56	1.12	-0.65	104	D104	0.86	44	0.20	0.39	1.02	0.05	0.86	0.33
94	D 94	0.96	0.14	0.37	2.07	69	D 69	0.64	1.30	-1.11	93	D 93	1.26	48	-1.07	0.42	1.03	0.07	0.84	0.28
95	D 95	-0.35	0.14	0.65	1.42	123	D123	0.68	1.09	-0.08	55	D 55	2.48	48	1.35	0.45	1.06	0.14	0.52	0.14
96	D 96	-1.58	0.19	1.72	-1.03	77	D 77	0.70	0.22	3.87	6	D 6	1.89	28	-0.44	0.46	1.04	0.10	0.68	0.23
97	D 97	0.74	0.14	0.68	1.21	35	D 35	0.70	1.03	-0.35	106	D106	-0.39	50	1.43	0.48	1.03	0.05	0.89	0.29
98	D 98	-3.53	0.41	1.09	0.02	102	D102	0.74	0.95	-0.15	89	D 89	0.90	5	-1.75	0.50	1.03	0.05	0.93	0.31
99	D 99	1.07	0.15	-0.72	6.08	97	D 97	0.74	0.68	1.21	42	D 42	-0.31	47	2.93	0.53	1.03	0.05	0.90	0.30
100	D100	-1.51	0.18	1.41	-0.46	107	D107	0.74	0.59	1.94	133	D133	1.59	29	0.05	0.61	1.05	0.08	0.85	0.27
101	D101	-0.27	0.14	0.76	1.09	67	D 67	0.74	0.26	2.94	118	D118	-0.21	49	-1.19	0.65	1.03	0.05	0.82	0.28

TABLE CONTINUED

SEQ NUM	ITEM NAME	ITEM DIFF	STD ERROR	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT TTEST	SEQ NUM	ITEM NAME	ITEM DIFF	INFO AMT	FIT BETWN	T-TESTS TOTAL	WTD MNSQ	MNSQ SD	DISC INDX	POINT BISER
102	D102	0.74	0.14	0.95	-0.15	21	D 21	0.76	1.44	-2.30	105	D105	1.15	50	0.46	0.66	1.04	0.06	0.84	0.27
103	D103	0.80	0.14	-0.64	6.81	103	D103	0.80	-0.64	6.81	116	D116	2.32	49	1.31	0.71	1.09	0.13	0.46	0.11
104	D104	0.86	0.14	0.86	0.39	104	D104	0.86	0.86	0.39	73	D 73	3.28	49	6.26	0.74	1.15	0.22	-0.64	-0.18
105	D105	1.15	0.15	0.84	0.66	89	D 89	0.90	0.93	0.50	130	D130	2.25	45	2.24	0.86	1.10	0.12	0.36	0.09
106	D106	-0.39	0.14	0.89	0.48	60	D 60	0.92	1.02	-0.17	137	D137	2.05	47	1.79	0.92	1.10	0.11	0.47	0.15
107	D107	0.74	0.14	0.59	1.94	94	D 94	0.96	0.37	2.07	74	D 74	-0.05	50	2.93	0.92	1.04	0.05	0.71	0.27
108	D108	0.06	0.14	1.35	-1.75	124	D124	0.98	-0.19	4.43	25	D 25	-1.29	51	3.12	0.96	1.10	0.10	0.54	0.14
109	D109	0.49	0.14	1.37	-1.76	50	D 50	0.98	0.59	1.54	127	D127	-0.09	51	0.42	0.97	1.04	0.05	0.85	0.28
110	D110	1.13	0.15	1.62	-2.43	117	D117	1.00	1.33	-1.44	129	D129	1.13	46	1.07	1.05	1.06	0.06	0.73	0.25
111	D111	0.39	0.14	1.73	-3.80	135	D135	1.00	1.45	-1.88	140	D140	1.98	51	1.41	1.08	1.11	0.10	0.53	0.13
112	D112	-0.27	0.14	1.16	-0.72	56	D 56	1.02	1.46	-2.17	101	D101	-0.27	49	1.76	1.09	1.05	0.05	0.76	0.26
113	D113	1.45	0.15	1.24	-0.77	99	D 99	1.07	-0.72	6.08	139	D139	1.95	41	1.43	1.12	1.11	0.10	0.47	0.13
114	D114	0.24	0.14	0.55	2.21	53	D 53	1.11	0.56	1.35	97	D 97	0.74	51	1.26	1.21	1.06	0.05	0.68	0.26
115	D115	0.41	0.14	1.11	-0.68	129	D129	1.13	0.73	1.05	134	D134	2.08	51	2.13	1.31	1.15	0.11	0.28	0.07
116	D116	2.32	0.19	0.46	0.71	110	D110	1.13	1.62	-2.43	53	D 53	1.11	26	1.30	1.35	1.08	0.06	0.56	0.24
117	D117	1.00	0.14	1.33	-1.44	105	D105	1.15	0.84	0.66	95	D 95	-0.35	47	1.27	1.42	1.08	0.05	0.65	0.22
118	D118	-0.21	0.14	0.82	0.65	93	D 93	1.26	0.84	0.42	14	D 14	1.65	49	2.16	1.50	1.13	0.08	0.40	0.13
119	D119	-1.73	0.20	0.79	0.16	62	D 62	1.28	1.22	-0.94	76	D 76	1.78	25	2.13	1.53	1.14	0.09	0.41	0.12
120	D120	-0.77	0.15	1.30	-0.82	11	D 11	1.40	0.74	-0.16	50	D 50	0.98	42	0.47	1.54	1.09	0.06	0.59	0.22
121	D121	0.18	0.14	1.12	-1.01	3	D 3	1.40	1.04	-0.31	83	D 83	-0.33	51	2.49	1.58	1.08	0.05	0.54	0.19
122	D122	0.20	0.14	1.26	-1.11	113	D113	1.45	1.24	-0.77	10	D 10	-0.86	51	2.57	1.81	1.14	0.07	0.17	0.09
123	D123	0.68	0.14	1.09	-0.08	44	D 44	1.47	1.06	0.24	107	D107	0.74	50	2.08	1.94	1.10	0.05	0.59	0.20
124	D124	0.98	0.14	-0.19	4.43	138	D138	1.47	1.22	-1.16	94	D 94	0.96	47	2.65	2.07	1.12	0.06	0.37	0.17
125	D125	1.78	0.17	1.08	-0.38	133	D133	1.59	0.85	0.61	61	D 61	-0.94	36	4.54	2.09	1.17	0.08	-0.09	0.01
126	D126	0.47	0.14	0.94	-0.14	14	D 14	1.65	0.40	1.50	114	D114	0.24	51	1.61	2.21	1.10	0.04	0.55	0.22
127	D127	-0.09	0.14	0.85	0.97	125	D125	1.78	1.08	-0.38	23	D 23	-0.33	50	2.16	2.32	1.12	0.05	0.42	0.14
128	D128	-0.37	0.14	1.56	-2.11	76	D 76	1.78	0.41	1.53	54	D 54	-0.13	48	2.50	2.33	1.11	0.05	0.55	0.19
129	D129	1.13	0.15	0.73	1.05	6	D 6	1.89	0.68	0.46	57	D 57	0.39	46	5.01	2.41	1.11	0.04	0.47	0.19
130	D130	2.25	0.19	0.36	0.86	139	D139	1.95	0.47	1.12	84	D 84	-0.63	28	5.09	2.78	1.18	0.06	-0.05	0.02
131	D131	-0.33	0.14	0.34	2.93	140	D140	1.98	0.53	1.08	131	D131	-0.33	48	2.73	2.93	1.16	0.05	0.34	0.12
132	D132	-1.12	0.17	1.57	-1.37	137	D137	2.05	0.47	0.92	67	D 67	0.74	36	2.73	2.94	1.15	0.05	0.26	0.15
133	D133	1.59	0.16	0.85	0.61	134	D134	2.08	0.28	1.31	46	D 46	0.22	39	3.64	3.42	1.15	0.04	0.34	0.14
134	D134	2.08	0.18	0.28	1.31	136	D136	2.08	0.89	0.18	88	D 88	-0.17	30	2.71	3.55	1.18	0.05	0.18	0.10
135	D135	1.00	0.14	1.45	-1.88	130	D130	2.25	0.36	0.86	87	D 87	-0.17	47	3.18	3.77	1.19	0.05	0.08	0.08
136	D136	2.08	0.18	0.89	0.18	116	D116	2.32	0.46	0.71	77	D 77	0.70	30	3.48	3.87	1.20	0.05	0.22	0.11
137	D137	2.05	0.18	0.47	0.92	55	D 55	2.48	0.52	0.45	124	D124	0.98	31	4.96	4.43	1.27	0.06	-0.19	0.00
138	D138	1.47	0.16	1.22	-1.16	38	D 38	2.79	0.57	0.07	99	D 99	1.07	41	7.71	6.08	1.40	0.06	-0.72	-0.17
139	D139	1.95	0.17	0.47	1.12	64	D 64	2.85	0.55	0.04	103	D103	0.80	33	6.89	6.81	1.39	0.05	-0.64	-0.14
140	D140	1.98	0.17	0.53	1.08	73	D 73	3.28	-0.64	0.74	85	D 85	0.49	32	8.73	8.91	1.45	0.04	-1.03	-0.24
	MEAN	0.00		1.04	-0.06										1.70	-0.06	0.99	0.08		
	S.D.	1.24		0.60	2.02										1.76	2.02	0.12	0.05		

140 ITEMS CALIBRATED ON 236 PERSONS
236 MEASURABLE PERSONS WITH MEAN ABILITY = 0.34 AND STD. DEV. = 0.74

7.2.4. Stability of Rasch estimates

The correlation between Rasch estimates of difficulty for items in Form A and those in either Form C or D is 0.93, while for those items in Form B and in either Form C or D it is 0.88.

7.2.5. Evaluation of Rasch statistics: measurement of fit

7.2.5.1. Measures available

The available measures for evaluating the statistics obtained by Rasch measurement were described in Chapter 5. These will now be briefly summarised before proceding to examine the results for this set of tests.

The 'analysis of fit' consists of a series of fit mean squares, which are mean square standardised residuals for item-by-person responses averaged over persons and partitioned into two components, one between ability groups and the other within ability groups. These mean squares increase in magnitude away from a reference value of 1 as the observed ICC departs from the expected ICC (see below) i.e. when too many high-ability persons fail an easy item or too many low-ability persons succeed on a difficult one. The statistical significance of large values can be judged by comparing the observed mean squares with their expected value of 1 in terms of the expected standard errors.

The *total* mean square evaluates the general agreement between the variable defined by the item and the variable defined by all other items over the whole sample.

The *between group* mean square evaluates the agreement between the observed ICC and the best fitting Rasch model curve over the ability sub-groups.

The *within-group* mean square summarises the degree of misfit remaining within ability groups after the *between group* misfit has been removed from the *total*.

The *discrimination index* given in this analysis describes the linear trend of departures from the model across ability groups expressed around a modal value of 1. When this index is near 1, then the observed and expected ICCs are close together over the reference points defined by the ability grouping.

Finally, an ICC analysis is available (not given here but available from the author) which gives the proportion of correct answers given by each ability group to each item; we expect the ICCs to increase as we move from left to right (from less able to

more able score groups). Large proportional departures from expected values indicate posssible misfit. In this analysis we use the between group mean square as an evaluator of ICCs. [A full description of the derivation of these statistics and their use can be found in Wright and Stone (1979) Chapter 4.]

7.2.5.2. Relationship between measures

To show just how varied these measures of fit can be, and to show how no one measure is to be relied upon (different purposes will require different measures), the correlation between values obtained for the various fit statistics are reported below. It will clearly be seen that reliance on one fit statistic only would be a serious mistake.

TABLE 5
CORRELATIONAL RELATIONSHIPS BETWEEN RASCH TEST STATISTICS

Form A

	Difficulty	Discrimination	Fit *t*-test
Difficulty	1.00	−0.36	0.11
Discrimination	–	1.00	−0.80
Fit *t*-test	–	–	1.00

Form B

Difficulty	1.00	−0.46	0.19
Discrimination	–	1.00	−0.92
Fit *t*-test	–	–	1.00

Form C

Difficulty	1.00	−0.37	0.12
Discrimination	–	1.00	−0.86
Fit *t*-test	–	–	1.00

Form D

Difficulty	1.00	−0.49	0.21
Discrimination	–	1.00	−0.89
Fit *t*-test	–	–	1.00

7.2.5.3. Fitting the items to the model

The best single measure for evaluating fit to the model is the fit *t* test, which is the basis for ordering the items in decreasing order of fit in the third panel of the tables just given. Looking at these tables reveals no consistent pattern of fit or misfit. It is not true, for example, that the harder items of Part 3 or the very easy items of

Part 2 are consistently worst fitting items, though individual forms of the test show different patterns.

In Form A it will be seen that Part 3 items either fit very well (18 items – nearly 50% – are found in the top third of the best-fitting items) or rather poorly (15 items – 39% – are found in the bottom 30%), though this may exaggerate the effect, since only at the extremes do the values of *t* depart markedly from the desired reference values. It will also be noted that the poorly fitting items tend also to have low point biserials (given in the final column). In fact, the correlation between the fit *t* test (total) and the value of the point biserial is 0.81 in Form A, 0..91 in Form B, 0.83 in Form C and 0.90 in Form D.

In Form A, then, there is no strong evidence for patterns of misfit. It is true that there are more or less well-fitting items, and that misfit can be interpreted in much the same way one would interpret classical statistics which show extreme values for facility values and discrimination indices. The question for Rasch modelling is whether such misfit is such as to disturb the estimation of the parameters. In Form A, only 17 items at the bottom end of the best-fit scale fall deviate by more than one standard deviation from the norm, which suggests that, in a test of 140 items, the model is working well. Moreover, 15 of those 17 items have already been identified as 'poor' by classical criteria. The source of the misfit is probably poor item writing, which can readily be corrected.

In Form B, a similar pattern (or lack of pattern) is seen; on this occasion there is a slightly greater proportion (19 – nearly 50%) of the Part 3 items in the bottom third of the scale, which suggests that this part of the test needs much revision. This had already been identified through classical analysis. Also there is a slightly greater number (20) of items which deviate by more than one standard deviation from the norm. We would probably have to say that Form B contains a greater proportion of misfit, for whatever reason, than Form A, though the source of the misfit is equally unclear.

Form C shows exactly the same pattern of fit as Form A, though the spread of fit is perhaps wider, suggesting that a wider range of ability is being considered. This is confirmed by the 'ability' summary given at the ends of the tables.

Form D shows a similar pattern to Forms A and C, though with a few extreme values at the bottom end of the scale for those items which were answered correctly by a large proportion of candidates.

The fit *t* test applied to these items as they occur in the test analysed a whole suggest, then, that very few of the items could be said not to fit the model used. Those that do not fit are probably poorly written items. Using criteria proposed by Wright, Mead and Bell (1980;84), who suggest an arbitrary upper limit for the fit *t* test of 2.00 (and which in fact is the limit set in this analysis for the identification of misfitting persons) we would have to reject 11 items in Form A, 22 items in Form B, 19 items in Form C, and 17 items in Form D. However, Wright, Mead and Bell also suggest (ib.) that in practice this value can be altered without any serious effect, depending on what inspection of the individual misfitting items reveals. In this case, inspection of the misfitting items, in all forms of the test, shows that the source of the problem is that items are simply too easy or too difficult for the sample. The 'easy' items will be discussed in the section on factor analysis. The 'difficult' items may be difficult either because they are genuinely difficult for this sample (in which case they cannot be rejected outright, merely reserved for those occasions when a high ability sample needs to be tested) or because they are poorly written. In the latter case, which is certainly what is happening with some of the items, re-writing is in order.

As a rough comparison with misfitting persons, it should be noted that in Form A 33 persons (12%) were deemed to be misfitting, in Form B there were 26 (11%), in Form C 25 (9%), and in Form D 30 (11%). This compares with the following percentages for misfitting items: Form A 8%, Form B 16%, Form C 14%, and Form D 12%. Clearly these figures are, allowing for the numbers involved, comparable. We conclude, therefore, that with minor reservations the Rasch model used here and the items tested fit very well.

7.2.6. Comparison of classical and Rasch statistics

There is a very high correlation between the facility values obtained in the various forms of the test and the Rasch estimates of difficulty (Form A r=0.99, Form B r=0.93, Form C r=0.90, and Form D r=0.98). This is not really surprising, since not only is the 'proportion correct' used as an initial estimate in the UCON procedure, but there is also a direct relationship between estimates of difficulty and proportion correct as was seen in chapter 5.

We have noted in the previous section the very high correlation between the fit *t* test and the point biserial. It was also noted that misfit often corresponds to extreme values of facility values and discrimination indices.

7.3. Dimensionality of the reading test items

7.3.1. Construct validity

7.3.1.1. Pearson correlation coefficients

As an initial check on the construct validity of the four forms of the test, the separate parts of the test (including Part 4, the summary-writing section) were correlated with each other. Results are as follows:

Table 6
Sub-test correlations

FORM A

	PART 1	PART 2	PART 3	PART 4	TOTAL
PART 1	1.00	0.80	0.65	0.42	0.92
PART 2	0.80	1.00	0.62	0.37	0.88
PART 3	0.65	0.62	1.00	0.29	0.79
PART 4	0.42	0.37	0.29	1.00	0.58
TOTAL	O.92	0.88	0.79	0.58	1.00

FORM B

PART 1	1.00	0.82	0.74	0.30	0.93
PART 2	0.82	1.00	0.72	0.35	0.92
PART 3	0.74	0.72	1.00	0.37	0.86
PART 4	0.30	O.35	0.37	1.00	0.51
TOTAL	0.93	0.92	0.86	0.51	1.00

FORM C

PART 1	1.00	0.84	0.83	0.38	0.94
PART 2	0.84	1.00	0.74	0.39	0.91
PART 3	0.83	0.74	1.00	0.36	0.89
PART 4	0.38	0.39	0.36	1.00	0.54
TOTAL	0.94	0.91	0.89	0.54	1.00

FORM D

PART 1	1.00	0.79	0.71	0.38	0.91
PART 2	0.79	1.00	0.70	0.35	0.89
PART 3	0.71	0.70	1.00	0.33	0.84
PART 4	0.38	0.35	0.33	1.00	0.57
TOTAL	0.91	0.89	0.84	0.57	1.00

It can easily be seen that Parts 1, 2, and 3 of the test in all forms correlate reasonably highly with each other, perhaps too highly – one might argue that the parts are not testing separate elements at all. Part 4, on the other hand, clearly shows such a low correlation with all the other parts that we should feel quite confident in claiming that it does indeed test something quite different (or that the reliability of that part of the test is too low).

Tentatively, then, we might expect on the basis of these figures that we have a test with two strands to it: a reading strand and a writing strand, but within the reading strand there does not appear to be the differentiation we had hoped to produce when we constructed the test. Further investigation of this aspect requires the use of factor analysis.

7.3.1.2. Factor analysis

As discussed in chapter 5, we use factor analysis in order to assess the dimensionality of a set of test items. In addition, factor analysis provides valuable information as to the construct validity of a test, and is therefore an important procedure in any test validation exercise.

The method of factor analysis used here is maximum likelihood with oblique rotation. The reasons for these choices depend partly on mathematical considerations and partly on considerations of *a priori* assumptions as to the structure of the data. Maximum likelihood was chosen simply because this appears to be the method most often used in other investigations of the validity of tests of this type (e.g. Lunzer *et al.* 1979, Spearrit 1972); in fact, as almost any handbook of factor analysis points out (e.g. Rummel 1970, Kim and Mueller 1978), the differences between results obtained by using different factoring methods are not usually great – often the choice is based upon mathematical elegance. Maximum likelihood has the additional advantage of being a procedure which is used in the estimation of various of the Rasch parameters and which therefore enables us to show consistency of analysis across the data.

In fact, various analyses using other factoring methods were tried on the data (Principal components, principal axis, alpha, and image factoring), but these are not reported here since the results were virtually indistinguishable from those obtained using the maximum likelihood method. The full analysis is available from the author.

More significant is the choice of rotation method. Oblique rotation was chosen for the important reason that in data of this kind (i.e. language test data of a reasonably homogeneous nature) one would *expect* factors to show some sort of relationship

with each other. It is unreasonable to suppose that factors obtained in a fairly narrowly defined test of reading will be orthogonal – an assumption which must be made if, for example, varimax rotation were used. This follows the argument used by Lunzer *et al.* 1979. The main problem which arises in using oblique rotation is that the structure of the factor matrix becomes more difficult to interpret. In the analysis which follows the structure matrix is reported where possible, sorted into groups with values below 0.3 suppressed to give a clearer picture of the data structure.

Again, for completeness, various other rotation methods were tried. These are also not reported since the results were essentially identical with those obtained using oblique rotation. We report the 'oblique' results because of our belief that they reflect more accurately the supposed structure of the data.

The data were analysed in various ways which are reported in the tables in the following pages. The whole test was analysed, each of the three parts of the test was analysed independently, and finally all combinations of two parts of the test were analysed together. Thus for each form of the test there are 7 analyses: Part 1, Part 2, Part3, Whole Test, Part 1 + Part 2, Part 2 + Part 3, and Part 1 + Part 3. This should give a comprehensive picture of the underlying structure of the test.

As an initial test of unidimensionality, Cattell's scree test was applied (see chapter 5 for a discussion of the logic behind this). The scree plots for Form A parts 1, 2, and 3 are reported here as an example of the kind of plots obtained from these data. No other plots are given here since they all, without exception, follow the same pattern as those shown here (full prints of scree plots are obtainable from the author). What these scree plots show is that by all normal criteria (and this is confirmed by the Reckase method discussed in chapter 5) all analyses of all parts or combinations of parts of the test in all forms result in one dominant factor, with other factors emerging at the 1.0 eigenvalue level hardly at all.

The importance of the scree test is that it shows just how arbitrary, and small, the extraction of factors round about the critical level (where the eigenvalue is 1.0) is. In other words, while it is possible to extract factors in some cases on the basis of their eigenvalues being greater than 1.0 (the usual criterion), in fact for the data reported here these extracted factors are virtually indistinguishable from the factorial litter which is found.

Here is strong evidence, then, for the unidimensionality of this set of items, in whatever combination they are taken. We can feel confident that all the test items used here form part of a single dimension, which we may choose to call 'reading in

English as a foreign language'. However, it is worth pursuing the analysis a little further in order to see if closer inspection reveals more subtle patterns within the data. It must still be borne in mind, though, that in everything which follows the overriding consideration must be that we are dealing with extremely fine tuning, and that in the overall scheme of things the test items exhibit, unequivocally, unidimensionality.

Analysis of Parts 1 and 3 of the test in all forms reveals no interesting patterns, and these are not reported here. However, this lack of pattern is itself interesting, since it shows quite clearly that there is no justification for dividing 'grammar' into separate subskills (which is what we would expect), in other words that there is no evidence that the ability to answer questions on, say, present/ perfect verb forms requires a different kind of ability to that required to answer, say, questions on prepositions. This is a fairly trivial observation in the context of a grammar test, but far less trivial in the case of 'reading comprehension' test. What this result in fact shows is that in reading English as a foreign language there is no evidence for the existence of identifiable sub-skills, at least as manifested in performance on a reading comprehension test. To that extent, this analysis shows for English as a foreign language what Lunzer *et al.* (1979) showed for English as a first language, namely that claims to identify reading subskills are unfounded.

Part 2 of the test, when analysed in isolation, does, however, reveal a little more. It will be seen that in Form A the group of items INT 12, 7 and 13 (and possibly 10) perform somehow differently from the other groups of items. (see Appendix I for the items and content attaching to these labels). In Form B it is items INT 4 and 13 which form one detached group and items INT 7, 9 and 5 which form another group. In Form C, this part of the test appears to fall into two evenly divided sets of items, one of which consists of INT 2, 1, 5, 3, 12, and 10. Finally in Part 2 of Form D, it is items INT 9 and 7 which form isolated groups, while INT 11, 8, 1, 10, 13 and 12 form one substantial group and the other items another. When these groups of items are matched to test content (remembering that the INT groups are designed to test the same content areas - and in some cases are the same items) it would appear that items relating to Information Transfer, Text Type Recognition and Logical Ordering (INT 12, 10, and 11 respectively) and possibly Connectors in Discourse (INT 13) somehow test a different underlying ability from the other items.

This is, of course, a highly tentative conclusion. Firstly, there is the problem that the extracted factors are not large. Secondly there is the difficulty of reconciling different patterns across forms of the test. However, there does appear to be slight

evidence for the existence of some sort of factor being present in items INT 10 – 13 which is not present in the other groups of items. This factor would appear to be either a 'test type' factor or a learning factor. It is difficult to distinguish between the two in this situation, since what appears to be happening is that students are taught how to do certain things with texts (e.g. state explicitly of what type a particular text is) which is an artificial teaching/testing exercise. This is not the sort of thing that occurs anywhere except in a teaching/testing environment for students of this kind, and therefore they can be taught how to deal with exercises/tests of this type, even though they are really nothing to do with the kind of reading activity they would normally be engaged in. Even information transfer, which one would have thought would have been a reasonably normal type of academic activity, is a taught and highly restricted exercise in this context, since the kind of text used to practise this 'skill' (the kind used in this test) is a highly artificial construction which bears little similarity with 'real-life' texts.

Analysis of Parts 1 and 3 together reveals a much clearer pattern (bearing in mind the caution that extracted factors are still vitually indistinguishable from scree beyond the first factor). Part 1, it will be remembered, could be called a test of grammar in a fairly restricted sense, while Part 3 could be called a test of reading in its more traditional sense of 'reading comprehension'. Here are the two extemes of our reading 'scale', if such a scale exists. Indeed, it will be seen that in all four forms of the test, the analysis of Parts 1 and 3 together shows quite clearly that the two parts are loaded on different factors. This result is not quite as clear as it might appear, however, since in Forms A, B and D there is a certain amount of mingling of item groups so that the loadings are never 'pure' in the sense that all the grammar items load exclusively on one factor and all the comprehension items load exclusively on another. Moreover, in Form C the analysis was unable to produce an oblique rotation of factors since only one factor was extracted, though the comprehension items tended to fall in the lower half of the factor loadings (this should not be taken as any sort of evidence, merely as a hopeful observation).

While not destroying our assumption of unidimensionality, then, it would be true to say that there appears to be some sort of factor associated with the grammar items and another factor associated with the comprehension items when these two sets of items are taken together.

When Parts 1 and 2 are analysed together, the only pattern which emerges is the pattern already described in relation to Part 2 in isolation. However, the situation is even less clear than it was there, and varies from form to form. In Form A (Parts 1

and 2) it will be seen that there are no obviously meaningful clusters of items, even though 5 separate factors can be extracted. In Form B (Parts 1 and 2) the items behave quite differently and only one meaningful factor can be extracted. Form C shows some of the Part 2 items loading together, though in the same way that they loaded in the analysis of Part 2 in isolation. For this group of testees there appears to be a difficulty associated with certain parts of the test which is probably attributable to the learning factor noted above (confirmed in this case by the fact that most of the identified 'learnable' groups of items load with the 'easier' grammar items and not with the more difficulty paragraph/paraphrase items). Form D (Parts 1 and 2) shows a similar pattern to Form A – an unclear set of factors comprising apparently unrelated items.

While it is difficult to see a consistent pattern in the analysis of Parts 1 and 2 together, what patterns exist appear to confirm the existence of a separate factor for certain items in Part 2. On the evidence of the data presented here it would be hard to argue that the step from simple grammar recognition items (largely single-word completion exercises) to paraphrase recognition items is a significant one.

Analysis of Parts 2 and 3 together suggests that there may be two separate factors involved, though it is far from clear what these might mean. In Form A it would be reasonable to suggest that on the evidence of the factor analysis Part 2 taps a different underlying ability from Part 3 (though there is still considerable contamination of the factor structure). Form B exhibits a similar, though more confused, pattern, as do Forms C and D. In fact, on closer inspection it would appear that the patterns relate closely to the patterns obtained in analysis of Part 2 in isolation. In other words the dominance of whatever feature it is in Part 2 that is causing clusters of items to be identified is strong enough to mask the strength of any factor which might be identifiable across parts of the test.

This is confirmed by the analysis of the whole test which shows, where it doesn't show the existence of one overriding factor as in Forms A and B, that the Part 2 items tend to group together (Form C) or group as 'learnable' items with a few items from both of the other parts of the test (Form D).

These results help to explain the rather odd results that were obtained in the classical item analysis (see section 7.1.1. above). If there is indeed some learning factor at work, then this could account for the unexpectedly high facility values observed in some of the Part 2 items; this could also explain differential performance, in that the same kind of Part 2 items discriminate very well in some forms of the test,

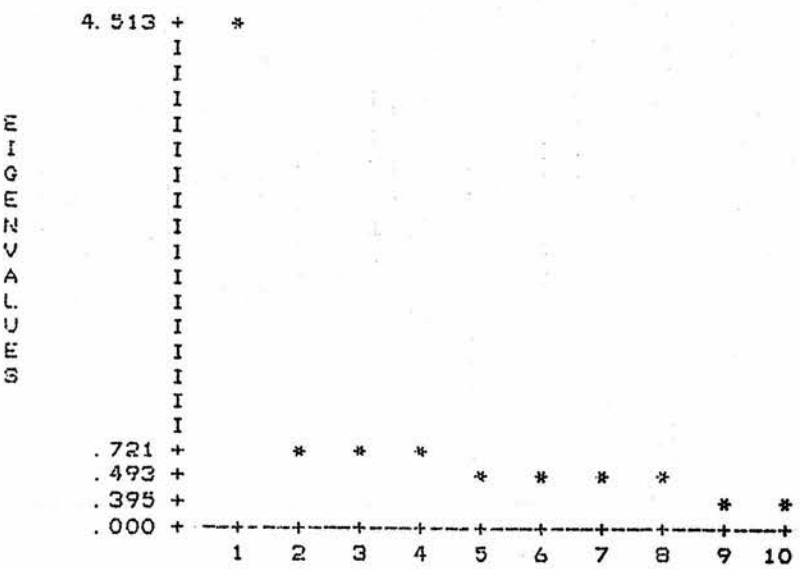
but not at all, or even negatively, in other forms.

In conclusion, there can be no doubt that the items analysed here are unidimensional however they are looked at. Observations as to the more detailed structure of the data must be largely negative: there do not appear to be easily identifiable 'subskills' for reading comprehension, nor do there appear to be easily identifiable components of reading if it is taken in a very broad sense. We might wish to say that some items are of such a type that they favour students who have either been schooled in test technique or been taught particular 'topics'. In addition, grammatical ability may be separable from ability in reading comprehension if these two abilities are tapped in the same test, though this may have more to do with stimulus material than with underlying ability. On the whole, there is little evidence to suggest that, for this group of students on this type of test, we should expect to find any differentiation in the ability which we have very broadly defined as 'reading in English as a foreign language'.

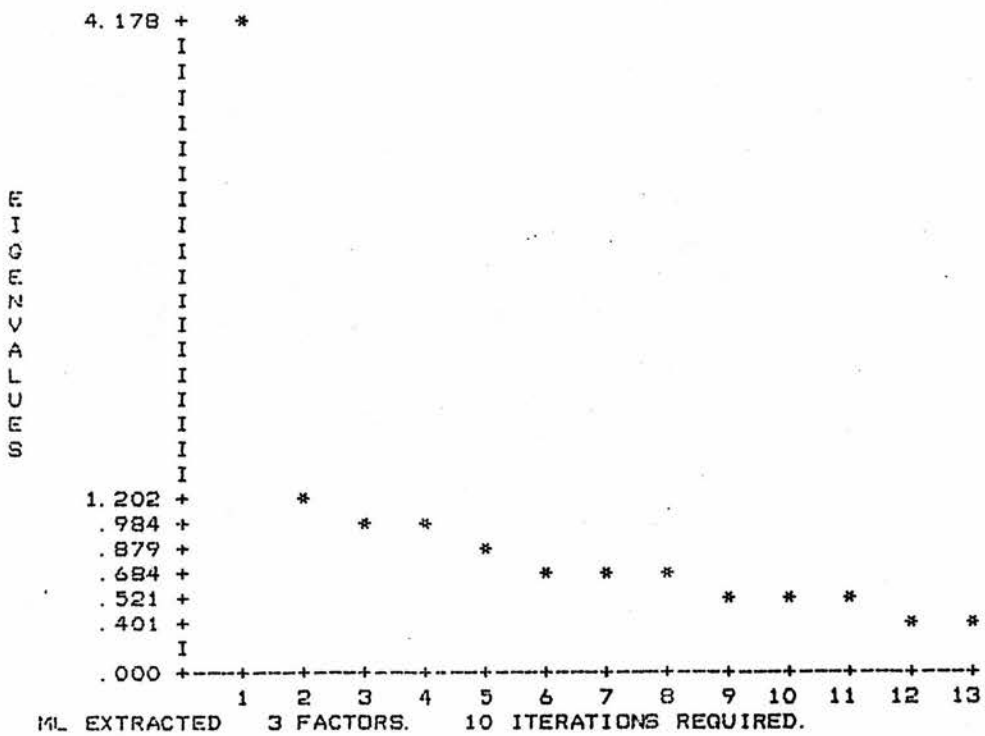
Factor Analysis of Form A

Scree Plots

1

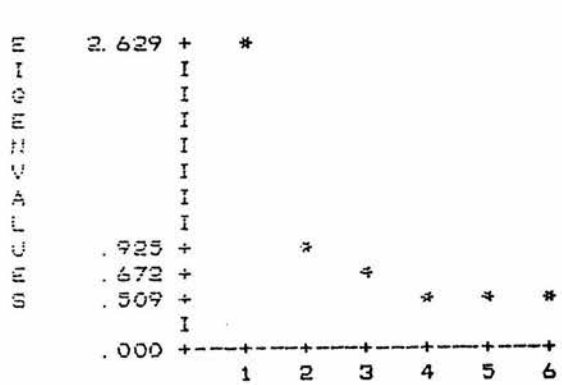


2



O ML EXTRACTED 3 FACTORS. 10 ITERATIONS REQUIRED.

3



FACTOR ANALYSIS OF FORM A

PART 2

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3
INT6	.69897		
INT5	.68631	-.47510	
INT4	.65974		
INT2	.60898	-.37023	
INT1	.58181		
INT3	.53869	-.46754	
INT8	.50101		
INT9	.43005		
INT11	.42348	-.31360	
INT10			
INT12		-.74185	
INT7			
INT13	.49655	-.53799	-.63164

IOLE TEST

OR MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5	FACTOR 6
1	.99948					
		.60606				
		.59772				
5	.33413	.58759				
9		.58312				
4	.33767	.57905				
		.56343				
8	.41899	.56168				
2		.54056				
	.37067	.53829	-.30651			
		.52865				
3		.49330				
		.48329	.37458			
		.47462				
7	.31297	.47269				
		.46966	.40093			
6		.45946				
1	.40566	.45186				
10	.39912	.45049				
	.33280	.44691				.37145
		.42549				
		.41251				
		.40048				
	.31717	.33197				
		.46583		.50457		
				.31810		

FACTOR ANALYSIS OF FORM A

ITS 1+2

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
INT11	.95910	.43760		-.35586	
SUBSC4		.72127		-.47493	.31105
SUBSC8		.70156	-.34236	-.53882	
SUBSC9		.64091		-.63561	
INT6		.62821		-.49108	.49372
INT4		.61435	-.32730	-.47010	.42893
INT1		.60677		-.39665	.34000
SUBSC10		.60279	-.37004	-.36895	
SUBSC7		.53517		-.43373	.34939
SUBSC6		.50605		-.35475	.32099
INT8		.49062	-.31318	-.36237	
INT9		.48529			
INT13		.42405	-.88964	-.33784	.33833
INT12			-.42079	-.35326	.33441
SUBSC2		.49862		-.79262	.31827
SUBSC5		.57292	-.37057	-.61784	.49416
SUBSC1		.50182		-.61773	.41978
SUBSC3		.50297	-.33612	-.55046	
INT3		.41491		-.50716	.49941
INT7					
INT5		.55611	-.42219	-.46287	.67861
INT2		.51084		-.48207	.57580
INT10					

ITS 1+3

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2
COMP5	.65789	-.41853
COMP2	.65467	-.39807
COMP7	.63309	-.42098
SUBSC4	.62441	-.60688
SUBSC10	.56695	-.49550
COMP3	.56151	-.44805
COMP6	.50644	-.31234
SUBSC6	.48025	-.42130
COMP4	.30416	
SUBSC2	.41964	-.72707
SUBSC9	.50613	-.68771
SUBSC5	.54005	-.67030
SUBSC1	.41622	-.65485
SUBSC8	.64620	-.65443
SUBSC3	.48584	-.57701
SUBSC7	.52162	-.52335

FACTOR ANALYSIS OF FORM A

PARTS 2+3

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
COMP2	.63874		-.35494	
COMP5	.63470		-.31197	.36869
COMP7	.61906		-.37694	
COMP3	.61699		-.42037	
INT1	.57658		-.53707	
COMP6	.57081		-.35061	
INT9	.52312		-.36252	
INT8	.48917		-.44672	
INT11	.48326	-.30268	-.38557	
INT13	.47237	-.43968	-.45613	
COMP4	.32609			
INT12		-.76291	-.31029	
INT6	.45588		-.72149	
INT5	.43607	-.40183	-.71779	
INT4	.52442		-.63357	.34828
INT2	.37343		-.62551	
INT3	.32361	-.35041	-.53726	
INT10				
INT7				.40721

FACTOR ANALYSIS OF FORM B

2T 2

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3
INT4	.96499		
INT13			
INT1		.55902	.38035
INT2		.53606	.35479
INT6		.52444	.52008
INT11		.43467	.31278
INT3		.39436	
INT8		.34574	
INT10		.31525	.30126
INT12			
INT7		.30002	.48596
INT9			
INT5			

FACTOR ANALYSIS OF FORM B

OLE TEST

'OR MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5	FACTOR 6
7	.99949					
C4		.81001				
C2		.77517				
C10		.73774	-.30832			
C8		.73049				
C1		.70872				
C9		.69646				
		.69286				
C6		.69256				
C3		.68681				
5		.66566				
C5		.65220				
C7		.63405				
2		.63031				
		.62861				
3		.61130				
		.58448				
		.55028				
3		.53115	.40770	.37222		
	.31383	.49347				
1		.48774		.30494		
		.46750				
4		.44237				
0		.41729				.31793
2		.40709				
6		.32376	.30757			
		.31894				

; 1+2

FACTOR MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
INT4	.99949			
SUBSC4	.32113	.78535		
SUBSC2		.76246		
SUBSC10		.74835		
SUBSC1		.72008		
INT2		.69643		
SUBSC8		.68827		
SUBSC3		.67417		
SUBSC6		.65385		
SUBSC7		.65020		
SUBSC9		.64671		
SUBSC5		.62997		
INT6		.61720		
INT3		.60091		
INT13		.58712		
INT1		.53159		
INT7		.49398		
INT11		.47918		
INT8		.47279		
INT10		.38666		
INT12		.37488		
INT9				
INT5				

FACTOR ANALYSIS OF FORM B.

ITS 1+3

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2
SUBSC4	.83437	.61333
SUBSC10	.78216	.41714
SUBSC2	.78120	.53162
SUBSC1	.76206	.52506
SUBSC8	.72744	.59344
SUBSC5	.69873	.44588
SUBSC6	.68830	.54191
SUBSC3	.67048	.60710
SUBSC7	.64882	.48863
COMP5	.64126	.73327
COMP2	.58661	.68681
SUBSC9	.67065	.67576
COMP3	.46327	.61408
COMP6		.49500
COMP4	.42078	.44407
COMP7		

S 2+3

RUCTURE MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
T2	.74223	.38233	.31224		
T6	.63970	.39511			.31683
T13	.59433	.54348			
T1	.54934	.33221	.36313		.32866
T3	.54733	.47935	.36062		
T10	.50868	.33361			
T7	.49753	.31420			
T11	.44674	.37394	.31204		.33995
T12	.40141	.33937			
T9	.31625				
MP3	.37289	.73820	.32088		.41218
MP5	.63107	.66882	.30587	-.40953	
MP2	.56390	.61343		-.35705	
MP4	.40048	.49839			
MP6		.48800			
T4			.67723		
T8	.48356	.52938		-.53254	
MP7				-.53023	
T5					.31845

FACTOR ANALYSIS OF FORM C

)RT 2

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2
INT6	.64536	-.35833
INT13	.61897	-.41350
INT4	.59557	-.43680
INT7	.55569	-.47535
INT8	.49147	
INT9	.46705	
INT11	.39322	-.37936
INT2	.56442	-.68830
INT1	.51972	-.66132
INT5	.42947	-.65436
INT3	.58521	-.60653
INT12		-.53947
INT10		-.41018

IOLE TEST

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3
SUBSC4	.76372	.48087	-.47168
SUBSC8	.75114	.49196	-.54143
SUBSC9	.74929	.37400	-.45113
COMP2	.74307	.37684	-.39698
COMP5	.72595	.46521	-.53839
SUBSC10	.67807	.49279	-.61691
COMP6	.66856	.40892	-.42808
SUBSC5	.64290	.58298	-.58375
INT6	.63092	.41003	-.46758
SUBSC3	.62256	.52843	-.31624
SUBSC1	.61514	.59877	-.43004
INT13	.60140	.42253	-.48140
SUBSC6	.59139	.47493	-.32197
COMP1	.59113	.35771	-.45661
COMP7	.58691	.32341	-.42190
COMP3	.56809	.37432	-.44675
SUBSC7	.54854	.37674	
INT8	.49977		
INT7	.49917	.46744	-.44054
INT9	.48648		
COMP4	.43678	.33040	-.32673
INT2	.58276	.70913	-.40575
INT1	.51020	.66003	-.47528
SUBSC2	.64485	.64510	-.39660
INT3	.54455	.63883	-.53640
INT5	.40439	.63132	-.36752
INT12		.53804	
INT11	.41530	.42587	
INT10		.37901	
INT4	.53223	.39528	-.77267

275 1.12

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2
SUBSC8	.76886	.43598
SUBSC9	.75266	
SUBSC4	.75047	.42174
SUBSC10	.71877	.46185
SUBSC5	.69559	.55266
SUBSC2	.66060	.58317
INT6	.64836	.35763
SUBSC1	.64338	.55152
INT13	.63010	.38183
SUBSC3	.61338	.46764
INT4	.58941	.42161
SUBSC6	.57395	.42881
SUBSC7	.53878	.31862
INT7	.52984	.44862
INT8	.49327	
INT9	.48830	
INT11	.43783	.37613
INT2	.61407	.67360
INT1	.54951	.65795
INT3	.59251	.62947
INT5	.44325	.62298
INT12		.53918
INT10		.39231

275 1.13

FACTOR MATRIX:

	FACTOR 1
SUBSC4	.76506
SUBSC8	.75347
COMP5	.74929
SUBSC9	.71916
COMP2	.71798
SUBSC10	.71262
SUBSC5	.69741
SUBSC2	.68576
COMP6	.66579
SUBSC1	.65225
SUBSC3	.63663
SUBSC6	.60862
COMP1	.60369
COMP7	.59179
COMP3	.58743
SUBSC7	.54623
COMP4	.45979

FACTOR ANALYSIS OF FORM C

PARTS 2+3

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2
COMP5	.75759	.48363
COMP2	.71338	.37209
COMP6	.68681	.41764
INT6	.63988	.40886
COMP7	.62825	.33684
COMP1	.62213	.37208
INT13	.60241	.45904
INT4	.59913	.47721
COMP3	.57746	.39667
INT7	.52477	.51410
INT8	.46812	
COMP4	.46433	.32715
INT9	.44770	
INT2	.57158	.70018
INT5	.42429	.67853
INT1	.53536	.67564
INT3	.57800	.63743
INT12		.52574
INT11	.39262	.40289
INT10		.37918

FACTOR ANALYSIS OF FORM D

PART 2

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
INT9	.99590			.31858
INT11		.61967	.31850	
INT8		.56314	.55028	
INT1		.56205	.41289	.47847
INT10		.51895	.30288	
INT13		.51269	.30788	
INT12		.38463		
INT3			.57800	
INT6		.44055	.54338	.54055
INT4			.53609	
INT2		.44009	.52743	
INT5		.31923	.40944	
INT7				.34521

FACTOR ANALYSIS OF FORM D

WHOLE TEST

FACTOR MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
SUBSC5	.82604	.56270			
SUBSC6	.82453	-.56491			
SUBSC2	.56756		.46820		
SUBSC10	.56250		.40367		
SUBSC3	.55031		.38526		
SUBSC1	.53578		.40508		
SUBSC9	.48380		.37175		
INT8	.46204		.43885		
SUBSC7	.44115		.39905		
INT2	.42684		.33037		
COMP3	.42411		.40920		
INT6	.42308		.38206		
COMP6	.42003		.41164		
INT4	.41827				
INT1	.41776		.39108		
INT9	.36406				
INT3	.36073		.32223		
INT10	.34596				
INT5	.33687				
INT7					
COMP4	.47376		.56913		
COMP2	.34147		.55826		
SUBSC4	.50909		.54785		
SUBSC8	.51792		.53311		
INT11	.32407		.41767		
COMP5	.36817		.39033		
INT13	.31047		.33732		
COMP7					
INT12					.32454

TS 1+2

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
INT9	.99089	.30566		.36993
SUBSC8		.79360	.53716	.36367
SUBSC4		.72094	.60409	.44097
SUBSC1		.65305	.57541	.53403
SUBSC5	.30126	.63491	.52368	.46575
SUBSC3		.62939	.48888	.58387
INT8		.59447	.54575	.37873
SUBSC9		.59237	.46606	.41465
SUBSC7		.57440	.54233	.30871
INT6		.57194	.40378	.34447
INT3		.53953		.33503
INT4		.53597		.42988
INT2		.52909	.38968	.34167
INT5		.44961	.30722	
SUBSC2		.65578	.67763	.51945
SUBSC10		.62127	.65945	.50647
INT11		.39922	.61904	
INT1		.48253	.54414	.34477
INT13		.37643	.53496	
INT10		.31454	.48691	
INT12			.33561	
SUBSC6		.48744	.37719	.66286
INT7				

FACTOR ANALYSIS OF FORM D.

PARTS 1 + 3

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2
SUBSC2	.74947	.60855
SUBSC4	.74129	.60773
SUBSC10	.73211	.50936
SUBSC8	.72739	.63895
SUBSC1	.72220	.47432
SUBSC5	.67309	.44679
SUBSC3	.66662	.57227
SUBSC7	.62390	.47205
SUBSC9	.61686	.51814
SUBSC6	.54609	.49474
COMP7		
COMP4	.60696	.79000
COMP2	.51889	.67138
COMP3	.48130	.65187
COMP5	.45349	.60858
COMP6	.54826	.56456

PARTS 2-3

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
COMP4	.74395	-.53671	.47102	.37754
COMP6	.63996	-.36838	.38578	
COMP3	.61503	-.46269	.36123	.30134
INT6	.57947		.43944	
INT3	.56453			
INT2	.53394	-.32121	.42704	
INT4	.52973	-.31441		
COMP5	.52188	-.41903	.32628	.39574
INT5	.42706	-.33767	.30052	
INT9	.37350			
INT7				
COMP2	.50316	-.99822	.38403	
INT11	.34977	-.34214	.58323	
INT10			.58024	
INT8	.53966	-.42456	.54996	
INT1	.46272	-.42735	.54527	
INT13	.34598		.50413	
COMP7				
INT12			.33137	.45196

7.3.2. Concurrent validity

As a measure of how well the test performed in relation to other available measures, student performance on the current test was correlated with student performance on the SPM1119 and SPM322 (school leaving certificate) tests. In addition the correlation between test results and current level of English as assigned by USM ('Class' in the following table) was calculated. Results were as follows:

TABLE 8
CRITERION CORRELATIONS

FORM A

	Part 1	Part 2	Part 3	Total	Class	SPM1119	SPM322
Part 1	1.00	0.79	0.66	0.92	0.71	0.75	0.83
Part 2	0.79	1.00	0.63	0.91	0.69	0.53	0.77
Part 3	0.66	0.63	1.00	0.84	0.60	0.60	0.65
Total	0.92	0.91	0.84	1.00	0.75	0.72	0.84
Class	0.71	0.69	0.60	0.75	1.00	0.59	0.82
SPM1119	0.75	0.53	0.60	0.72	0.59	1.00	0.73
SPM322	0.83	0.77	0.65	0.84	0.82	0.73	1.00

FORM B

Part 1	1.00	0.82	0.73	0.95	0.69	0.77	0.85
Part 2	0.82	1.00	0.71	0.93	0.63	0.70	0.79
Part 3	0.73	0.71	1.00	0.86	0.60	0.76	0.69
Total	0.95	0.93	0.86	1.00	0.71	0.80	0.86
Class	0.69	0.63	0.60	0.71	1.00	0.72	0.74
SPM1119	0.77	0.70	0.76	0.80	0.72	1.00	0.70
SMP2	0.85	0.79	0.69	0.86	0.74	0.70	1.00

FORM C

Part 1	1.00	0.83	0.82	0.96	0.78	0.76	0.86
Part 2	0.83	1.00	0.73	0.92	0.68	0.67	0.76
Part 3	0.82	0.73	1.00	0.91	0.75	0.71	0.76
Total	0.96	0.92	0.91	1.00	0.79	0.76	0.86
Class	0.78	0.68	0.75	0.79	1.00	0.66	0.81
SPM1119	0.76	0.67	0.71	0.76	0.66	1.00	0.04
SPM322	0.86	0.76	0.76	0.86	0.81	0.04	1.00

FORM D

Part 1	1.00	0.78	0.71	0.93	0.72	0.71	0.84
Part 2	0.78	1.00	0.70	0.92	0.70	0.66	0.79
Part 3	0.71	0.70	1.00	0.87	0.61	0.74	0.72
Total	0.93	0.92	0.87	1.00	0.75	0.77	0.86

Class	0.72	0.70	0.61	0.75	1.00	0.71	0.83
SPM1119	0.71	0.66	0.74	0.77	0.71	1.00	0.58
SPM322	0.84	0.79	0.72	0.86	0.83	0.58	1.00

There is a slight problem here (as with all criterion-related validities) in that if the test shows a high correlation with already available measures it may be questioned whether there is any need for a new test at all. Conversely, if the test shows a low correlation with those measures, then one would be inclined to think that the test is not performing as well as what is already available. Within these limitations, it will be seen that the test shows moderate to high correlations both with previous test performance and with present class standing. These figures are capable of supporting the argument that the test is sufficiently different from what is already available to justify its continued existence. They are also capable of supporting the argument that the test appears to be doing something so different that we should be extremely cautious in proceding. The argument is not resolvable here and is essentially a policy decision. What the figures do not show is whether the test shows a difference because it is a good test or because it is a bad test; this again is an argument that must be resolved elsewhere. Criterion correlations, while interesting, can do no more than provide a fraction of the evidence needed in consideration of the broader issues and will not be discussed further here.

7.4. Difficulty: using Rasch analysis with items

7.4.1. Deriving and comparing estimates of difficulty

The difficulty parameter estimates for the items of all forms were given in section 7.1, where it was assumed that all 140 items could be calibrated on a common scale. The rationale for this was based on considerations of content and test design as outlined in chapter 6 and also on considerations of dimensionality, investigated in section 7.2.

However, it is also possible to derive estimates of difficulty by considering the test in three component parts i.e. by forcing upon it the assumption that the three parts really form separate dimensions and should therefore be treated separately. If the different parts of the test really do constitute separate dimensions of the reading construct, then estimates of difficulty derived by considering the parts in isolation should be quite different from those derived by considering the test as a whole.

The tables reported here show the estimates of difficulty which are obtained when

Table 8a

DIFFICULTY ESTIMATES (BY PART TEST)

SEQ NUM	ITEM NAME	ITEM DIFF
1	A 1	0.34
2	A 2	-1.74
3	A 3	-1.29
4	A 4	-0.59
5	A 5	1.58
6	A 6	-0.71
7	A 7	-1.80
8	A 8	0.49
9	A 9	-1.47
10	A 10	-1.47
11	A 11	-1.92
12	A 12	-0.45
13	A 13	0.59
14	A 14	1.60
15	A 15	3.29
16	A 16	-1.52
17	A 17	1.13
18	A 18	0.30
19	A 19	-1.16
20	A 20	-1.29
21	A 21	-3.26
22	A 22	0.53
23	A 23	-0.37
24	A 24	-2.23
25	A 25	1.42
26	A 26	0.84
27	A 27	0.61
28	A 28	-0.68
29	A 29	-1.74
30	A 30	1.54
31	A 31	0.17
32	A 32	-1.99
33	A 33	-0.87
34	A 34	2.27
35	A 35	-1.42
36	A 36	-0.42
37	A 37	1.36
38	A 38	2.04
39	A 39	1.60
40	A 40	0.34
41	A 41	3.81
42	A 42	-0.17
43	A 43	0.32
44	A 44	-1.74
45	A 45	-1.74
46	A 46	1.50
47	A 47	1.83
48	A 48	0.90
49	A 49	-1.99
50	A 50	3.60

ITEM NAME	ITEM DIFF
A 51	0.73
A 52	-0.89
A 53	-1.77
A 54	-0.03
A 55	1.31
A 56	-1.41
A 57	1.41
A 58	-1.35
A 59	3.81
A 60	1.24
A 61	0.54
A 62	-1.69
A 63	1.73
A 64	-0.58
A 65	2.91
A 66	-0.47
A 67	-0.34
A 68	-0.76
A 69	1.20
A 70	1.59
A 71	0.06
A 72	-0.98
A 73	0.85
A 74	-0.08
A 75	0.35
A 76	4.35
A 77	3.48
A 78	-0.38
A 79	0.28
A 80	0.28
A 81	0.33
A 82	-0.61
A 83	0.44
A 84	-0.14
A 85	-0.89
A 86	-0.73
A 87	-0.41
A 88	-1.29
A 89	-2.59
A 90	-2.29
A 91	-2.17
A 92	-1.95
A 93	-2.59
A 95	0.35
A 96	-1.61
A 97	1.42
A 98	-2.43
A 99	2.36
A100	-0.58

ITEM NAME	ITEM DIFF
A101	-0.95
A102	0.57
A103	0.57
A104	-1.77
A105	-0.87
A106	-0.95
A107	0.55
A108	-0.54
A109	1.97
A110	-0.34
A111	0.40
A112	0.33
A113	-1.22
A114	0.48
A115	-0.87
A116	0.55
A117	-0.38
A118	-2.15
A119	-1.07
A120	-1.63
A121	0.67
A122	-2.39
A123	-0.34
A124	0.40
A125	0.49
A126	-0.50
A127	0.19
A128	-0.50
A129	1.56
A130	-0.17
A131	0.69
A132	0.49
A133	1.17
A134	0.95
A135	1.47
A136	-0.34
A137	1.80
A138	-0.32
A139	0.59
A140	1.22

SEQ NUM	ITEM NAME	ITEM DIFF
1	B 1	0.42
2	B 2	2.21
3	B 3	1.08
4	B 4	-0.18
5	B 5	-0.39
6	B 6	1.82
7	B 7	-1.04
8	B 8	1.33
9	B 9	0.16
10	B 10	0.02
11	B 11	-1.04
12	B 12	-0.45
13	B 13	0.72
14	B 14	0.87
15	B 15	3.53
16	B 16	-0.74
17	B 17	-0.39
18	B 18	-0.16
19	B 19	-1.15
20	B 20	-0.53
21	B 21	1.19
22	B 22	-0.56
23	B 23	0.94
24	B 24	-0.77
25	B 25	1.44
26	B 26	-1.18
27	B 27	1.10
28	B 28	-0.23
29	B 29	0.63
30	B 30	-0.45
31	B 31	-0.68
32	B 32	2.44
33	B 33	-0.06
34	B 34	1.21
35	B 35	-0.59
36	B 36	-1.26
37	B 37	0.54
38	B 38	-2.43
39	B 39	0.58
40	B 40	-1.76
41	B 41	-1.00
42	B 42	-1.66
43	B 43	0.42
44	B 44	-3.10
45	B 45	-0.97
46	B 46	0.54
47	B 47	2.00
48	B 48	-0.50
49	B 49	-0.62
50	B 50	-1.30

ITEM NAME	ITEM DIFF
B 51	1.21
B 52	0.10
B 53	0.99
B 54	0.27
B 55	-0.03
B 56	1.10
B 57	0.18
B 58	-0.58
B 59	0.08
B 60	4.65
B 61	0.82
B 62	-0.58
B 63	3.04
B 64	-0.41
B 65	-2.70
B 66	-1.04
B 67	0.29
B 68	-0.25
B 69	-1.20
B 70	-0.89
B 71	1.76
B 72	-1.52
B 73	-0.92
B 74	-0.11
B 75	-0.76
B 76	-0.01
B 77	1.41
B 78	-3.33
B 79	-1.07
B 80	-0.98
B 81	-0.38
B 82	-0.60
B 83	-0.25
B 84	-1.10
B 85	0.84
B 86	-0.18
B 87	0.06
B 88	0.08
B 89	1.08
B 90	-0.11
B 91	-0.48
B 92	-0.55
B 93	1.68
B 94	1.63
B 95	0.49
B 96	0.66
B 97	0.25
B 98	-0.20
B 99	-0.76
B100	-1.69

ITEM NAME	ITEM DIFF
B101	0.28
B102	-0.60
B103	-1.96
B104	-1.50
B105	-0.35
B106	-0.68
B107	0.05
B108	0.46
B109	-0.70
B110	0.24
B111	0.53
B112	0.32
B113	0.34
B114	-1.04
B115	-0.08
B116	-0.39
B117	0.61
B118	0.13
B119	-3.12
B120	-0.60
B121	1.12
B122	-0.31
B123	-0.66
B124	-0.10
B125	-0.70
B126	0.13
B127	0.53
B128	1.01
B129	0.86
B130	-0.31
B131	0.26
B132	-0.42
B133	0.44
B134	1.54
B135	0.32
B136	1.06
B137	0.28
B138	0.75
B139	1.35
B140	0.89

SEQ NUM	ITEM NAME	ITEM DIFF
1	C 1	0.31
2	C 2	-1.04
3	C 3	0.52
4	C 4	-0.43
5	C 5	-0.81
6	C 6	0.14
7	C 7	-1.45
8	C 8	1.24
9	C 9	1.06
10	C 10	0.86
11	C 11	-2.13
12	C 12	1.08
13	C 13	0.46
14	C 14	0.96
15	C 15	1.08
16	C 16	-0.02
17	C 17	-0.96
18	C 18	-1.90
19	C 19	0.14
20	C 20	1.85
21	C 21	-2.03
22	C 22	1.58
23	C 23	-1.55
24	C 24	-1.94
25	C 25	1.30
26	C 26	-0.89
27	C 27	3.18
28	C 28	-0.79
29	C 29	-1.04
30	C 30	0.71
31	C 31	1.92
32	C 32	-0.45
33	C 33	0.16
34	C 34	0.52
35	C 35	-1.12
36	C 36	-1.15
37	C 37	-3.05
38	C 38	-0.16
39	C 39	1.58
40	C 40	1.30
41	C 41	0.62
42	C 42	-0.51
43	C 43	-0.84
44	C 44	1.26
45	C 45	0.43
46	C 46	1.41
47	C 47	-0.86
48	C 48	-1.42
49	C 49	1.06
50	C 50	-0.20

ITEM NAME	ITEM DIFF
C 51	1.66
C 52	-1.19
C 53	0.67
C 54	-0.45
C 55	1.07
C 56	0.29
C 57	1.83
C 58	0.78
C 59	1.33
C 60	0.35
C 61	0.39
C 62	0.91
C 63	0.11
C 64	-1.15
C 65	0.37
C 66	-1.11
C 67	-0.56
C 68	-0.86
C 69	5.15
C 70	-1.74
C 71	-0.18
C 72	0.82
C 73	1.66
C 74	1.64
C 75	0.82
C 76	0.87
C 77	3.27
C 78	-0.15
C 79	0.39
C 80	0.50
C 81	0.62
C 82	-1.32
C 83	0.56
C 84	0.74
C 85	-1.11
C 86	-0.62
C 87	-0.45
C 88	-2.08
C 89	-3.00
C 90	-2.57
C 91	-2.16
C 92	-3.00
C 93	-3.19
C 94	-3.42
C 95	1.08
C 96	1.28
C 97	1.30
C 98	-0.22
C 99	0.78
C100	-0.70

ITEM NAME	ITEM DIFF
C101	-1.24
C102	-2.02
C103	0.11
C104	-2.49
C105	0.31
C106	0.40
C107	0.12
C108	0.58
C109	-0.77
C110	0.75
C111	1.01
C112	-1.51
C113	0.19
C114	-0.28
C115	0.94
C116	0.31
C117	2.18
C118	0.23
C119	0.51
C120	0.49
C121	-0.83
C122	0.96
C123	-0.74
C124	0.81
C125	-0.10
C126	-1.35
C127	-0.42
C128	-0.77
C129	1.03
C130	-1.44
C131	0.23
C132	0.92
C133	-0.94
C134	0.60
C135	1.75
C136	0.62
C137	1.59
C138	-0.37
C139	-0.42
C140	-0.92

DIFFICULTY ESTIMATES (BY PART TEST)

SEQ NUM	ITEM NAME	ITEM DIFF
1	D 1	-1.43
2	D 2	-0.85
3	D 3	1.99
4	D 4	0.12
5	D 5	0.55
6	D 6	2.42
7	D 7	-1.10
8	D 8	-0.90
9	D 9	-1.92
10	D 10	-0.55
11	D 11	1.87
12	D 12	0.71
13	D 13	-0.55
14	D 14	2.20
15	D 15	-0.98
16	D 16	-1.01
17	D 17	-0.59
18	D 18	-1.26
19	D 19	-1.88
20	D 20	-0.96
21	D 21	1.32
22	D 22	-2.08
23	D 23	0.04
24	D 24	0.26
25	D 25	-0.87
26	D 26	0.44
27	D 27	-0.25
28	D 28	-0.80
29	D 29	-0.34
30	D 30	0.46
31	D 31	-0.69
32	D 32	0.59
33	D 33	-1.26
34	D 34	-1.22
35	D 35	1.20
36	D 36	0.92
37	D 37	-0.04
38	D 38	3.64
39	D 39	-0.12
40	D 40	-0.55
41	D 41	0.79
42	D 42	0.00
43	D 43	-0.41
44	D 44	2.09
45	D 45	0.50
46	D 46	0.73
47	D 47	-0.27
48	D 48	-1.29
49	D 49	-0.23
50	D 50	1.58

ITEM NAME	ITEM DIFF
D 51	-1.56
D 52	-0.83
D 53	1.16
D 54	-0.13
D 55	2.44
D 56	1.06
D 57	0.47
D 58	0.56
D 59	-0.97
D 60	0.94
D 61	-0.97
D 62	1.33
D 63	-1.66
D 64	2.74
D 65	-1.09
D 66	0.16
D 67	0.88
D 68	-1.26
D 69	0.67
D 70	-0.97
D 71	-0.36
D 72	-0.85
D 73	3.07
D 74	-0.02
D 75	-0.40
D 76	1.78
D 77	0.66
D 78	-0.54
D 79	-0.69
D 80	-0.40
D 81	-1.09
D 82	-1.17
D 83	-0.27
D 84	-0.61
D 85	0.45
D 86	-0.28
D 87	-0.15
D 88	-0.15
D 89	0.88
D 90	0.00
D 91	-0.34
D 92	-0.36
D 93	1.31
D 94	1.00
D 95	-0.27
D 96	-1.53
D 97	0.79
D 98	-2.97
D 99	1.06
D100	-1.49

ITEM NAME	ITEM DIFF
D101	-0.96
D102	0.01
D103	-0.02
D104	0.16
D105	0.37
D106	-1.17
D107	-0.01
D108	-0.58
D109	-0.17
D110	0.44
D111	-0.22
D112	-0.89
D113	0.76
D114	-0.36
D115	-0.29
D116	1.42
D117	0.29
D118	-0.82
D119	-2.23
D120	-1.28
D121	-0.55
D122	-0.49
D123	-0.06
D124	0.27
D125	0.97
D126	-0.20
D127	-0.84
D128	-0.98
D129	0.40
D130	1.45
D131	-1.05
D132	-1.50
D133	0.90
D134	1.18
D135	0.33
D136	1.39
D137	1.30
D138	0.76
D139	1.21
D140	1.08

each of the three parts of the test are treated in isolation. The full statistics for this analysis are available from the author, but are not reported here since they do not materially affect the argument, nor do they show any real differences from patterns already reported in section 7.1. It might possibly be argued that other constructs could be imposed on the test structure; that Parts 1 and 2 should be treated together, for example, and Part 3 treated in isolation, much as we did in the factor analysis of section 7.2. In fact these analyses were carried out and are again available from the author, but the results are of no added significance and are not reported here.

It is obvious from a glance at the figures given here and a comparison with the figures given in section 7.1 that the difficulty estimates are different (e.g. item A1 in the whole test analysis has a difficulty of 0.24, while in the individual part analysis it has a difficulty of 0.34). If this difference is a purely linear relationship, then there is no problem in reconciling the two values. If, however, the relationship is not linear, then we have a real problem, since we are committed to saying that estimates of difficulty are essentially dependent on the items with which they are calibrated, and we are in danger of losing the benefits of sample-freeness.

This is the first step in validating the items for item banking purposes, and it is essential that we investigate this relationship, otherwise we may proceed no further. If stable estimates cannot be produced from the same group of testees, there would be little point in trying to obtain stable parameter estimates from different groups of testees.

Now, one might expect an analysis of this kind automatically to produce estimates related in a non-linear fashion. The reason for this would be that the Rasch procedure in essence normalises distributions and corrects for anomalies in the sampling groups. Thus one would expect a set of very 'easy' or very 'difficult' items to be spread so that the ends of the distribution were more usable. So if a set of difficult items (as our Part 3 items tend to be) are calibrated by themselves (as we do in this section) we would expect some of them to be given perhaps artificially lower values (i.e. they would be made 'easier') than if they are calibrated as part of a larger scale, where they could quite happily occupy the 'difficult' end of the scale.

As a measure of the relationship between estimates obtained in the whole test analysis of section 7.1 and the individual part analysis given here, correlation coefficients were calculated and were as follows:

Table 9
Relationship between Rasch difficulty estimates derived from the whole test /cont.

Table 9
Relationship between Rasch difficulty estimates derived from the whole test with those derived from sub-tests

	Part 1	Part 2	Part 3
FORM A			
Whole test	0.997	0.911	0.998
FORM B			
Whole test	0.986	0.998	0.999
FORM C			
Whole test	0.999	0.995	0.999
FORM D			
Whole test	0.980	0.993	0.997

It will be seen that there is a very high correlation between the estimates obtained in whole-test and in part-test analysis. We are therefore justified in claiming a strong linear relationship between estimates obtained thus. Practically, this means that we may use either analysis, as is convenient, for the derivation of parameter estimates.

Why, then, do we not find the expected non-linear relationship? The most reasonable explanation is that the unidimensionality of the whole set of items is strong enough to offset any deviations which might occur by isolating smaller subsets of items. This provides additional evidence for the unidimensionality of the item pool, and indeed strengthens that evidence, since even by pushing the assumption of multi-dimensionality to its limit (which is what we are in effect doing by analysing the test in this way) we achieve nothing that we do not achieve by assuming unidimensionality.

7.4.2. Difficulty of texts and tasks: Parts 1 and 2

This section and the next digress slightly to look at the content of the items in relation to their difficulty. We are concerned with answering the question: what is difficult/easy about these items?

For Parts 1 and 2, the items were grouped according to content area (see Appendix I) and the average difficulty value for items in each area was calculated. The calculation was performed twice: once for each part considered by itself and once in combination with the other part. Results (in increasing order of difficulty) were as follows:

Table 10
Rank order of average difficulty of items in Parts 1 and 2

	All forms			Form A			Form B			Form C			Form D		
	Pt 1	Pt 2	1+2	Pt 1	Pt 2	1+2	Pt 1	Pt 2	1+2	Pt 1	Pt 2	1+2	Pt 1	Pt 2	1+2
1	CP	SD	SD	PP	IT	IT	CP	SD	CP	CJ	IT	IT	CP	SD	CP
2	AR	PR	PR	CJ	SD	SD	AR	LO	SD	AR	SD	SD	AR	LO	SD
3	PP	IT	CP	AR	PR	PR	PF	AP	LO	PP	PR	PR	PR	CD	LO
4	CJ	LO	IT	PF	SE	SE	PP	CE	AR	PR	AP	AP	IG	PR	AR
5	PF	AP	LO	CP	LO	LO	FJ	C	AP	CP	CE	CJ	PP	AP	PE
6	PE	CE	AR	PE	CE	PP	PE	PR	PF	MA	LO	CE	PF	PC	IG
7	IG	CD	PP	IG	RS	CE	IG	CD	CE	CO	RC	AR	CJ	RS	CD
8	MA	SE	CJ	RP	CD	RS	MA	TT	C	PE	CD	PP	MA	SE	PP
9	RP	C	AP	MA	C	CJ	FO	SE	PP	RP	SE	LO	RP	C	PF
10	CO	RS	PF	CO	AP	CD	RP	IT	CJ	IG	C	PF	CO	IT	PR
11	–	TT	CE	–	PC	C	–	RC	PE	–	RS	RC	–	TT	CJ
12	–	PC	CD	–	TT	AR	–	PC	PR	–	TT	CD	–	CE	AR
13	–	RC	OE	–	RC	OF	–	RS	CD	–	OC	CM	–	RC	PC
14	–	–	SE	–	–	CM	–	–	TT	–	–	SE	–	–	RS
15	–	–	C	–	–	PE	–	–	IG	–	–	CO	–	–	MA
16	–	–	IG	–	–	IG	–	–	MA	–	–	MA	–	–	SE
17	–	–	MA	–	–	AP	–	–	SE	–	–	PE	–	–	C
18	–	–	RP	–	–	RP	–	–	IT	–	–	RP	–	–	RP
19	–	–	CO	–	–	MA	–	–	CO	–	–	C	–	–	IT
20	–	–	RS	–	–	CO	–	–	RP	–	–	IG	–	–	TT
21	–	–	TT	–	–	PC	–	–	RC	–	–	RS	–	–	CO
22	–	–	PC	–	–	TT	–	–	PC	–	–	TT	–	–	CE
23	–	–	RC	–	–	RC	–	–	RSD	–	–	PC	–	–	RC

Table 11
Spearman Rank Order Correlations for difficulties of Part 1 and Part 2 items

	Part 1	Part 2	Parts 1+2
$\rho_{Total - A}$	0.85	0.94	0.74
$\rho_{Total - B}$	0.96	0.70	0.73

$\rho_{Total - C}$	0.64	0.83	0.78
$\rho_{Total - D}$	0.77	0.58	0.56

Clearly a consistent relationship holds between the orders of difficulty obtained here. What is less clear is the extent to which we should say that such-and-such an item type is intrinsically 'easier' than another. The interpretation of orders of this type is complex (see Hill 1984 for a discussion of this issue in connection with 'natural orders' of acquisition), and following the methods used in morpheme acquisition studies we could probably group items into clusters, though what these orders mean in any easily interpretable sense we would find difficult to say. However, the following clusters appear to be identifiable:

Table 12
Clusters of items for Part 1 and Part 2

Part 1

CP+AR+PP → CJ+PF+PE → MA+RP+CD+IG

Part 2

SD+PR+LO+IT → AP+CE+CN+SE → RS+PC+RC+CO+TT

Parts 1+2

SD+PR+CP+LO+IT → AR+PP+CJ → AP+PF+CE+CN+PE → SE+CO+IG+MA+RP → CD+RS+TT+PC+RC

(KEY: CP=COMPARISONS AR=ARTICLES PP=PAST/PERFECT CJ=CONJUNCTIONS PF=PRESENT/FUTURE PE=PREPOSITIONS MA=MODALS/AUXILIARIES RP=RELATIVE PRONOUNS CD=CONDITIONALS IG=INFINITIVES/GERUNDS

SD=SIMILARITY/DIFFERENCE PR=PURPOSE/RESULT LO=LOGICAL ORDERING IT=INFORMATION TRANSFER AP=ACTIVE/PASSIVE CE=CAUSE/EFFECT CN=CONNECTORS IN DISCOURSE SE=SEQUENCE OF EVENTS RS=REPORTED SPEECH PC=POSSIBILITY/CERTAINTY RC=RELATIVE CLAUSES CO=CONNECTORS TT=TEXT TYPES)

The fact that items from parts 1 and 2 are well mixed up here suggests that there is no rigid distinction between them and that in this context they could be said to be ranged on the same scale.

7.4.3. Difficulty of texts and tasks: Part 3

The procedure outlined at the beginning of the previous section was repeated for the items in Part 3, taking each comprehension passage as a single area of content. This differs from the previous analysis in that the unit of content is not that defined by the item but that defined by the comprehension passage. The reason for this is that in analysis of comprehension items, all items belonging to a single passage must be treated together (as discussed in chapter 1), and that there is no logical basis for grouping items across passages because of the requirement (in classical as much as in IRT theory) of item independence. Unfortunately, the results obtained will not be very interesting from a content point of view. The results obtained (including texts used at the pre-pilot stage) were as follows:

Table 13
Text and task difficulty for Part 3

	Text title	Difficulty	Flesch (RE) Index
1	Digital computers	-3.34	8.26
2	Noise	-1.39	45.37
3	Nurse Maitland	-1.09	48.97
4	Chelation	-1.09	53.59
5	Ethical dimensions	-0.97	53.98
6	Modern surgery	-0.91	48.26
7	Insecticides	-0.91	57.55
8	Social evolution	-0.91	34.04
9	Electric fish	-0.51	58.62
10	Rabies	-0.34	36.69
11	Motorcycles	-0.26	75.93
12	Green banana	-0.26	69.20
13	Metal bending	-0.15	61.41
14	Experimenter effect	-0.05	14.22
15	Universities	0.17	33.11
16	Animals and language	0.43	27.51
17	Smoking	0.64	33.90
18	Kizzy	0.64	80.28
19	Automation	0.72	36.16
20	Dilemmas	0.77	66.97
21	Technology	1.86	29.22
22	Noise ...	2.15	21.01
23	Popular expositions	2.34	21.01
24	Int. Technology	2.42	56.52

This table gives us an idea of how the passages stand in relationship to each other (or rather how the passages plus items relate), but little else. Questions about item content have been answered in section 7.2 (factor analysis).

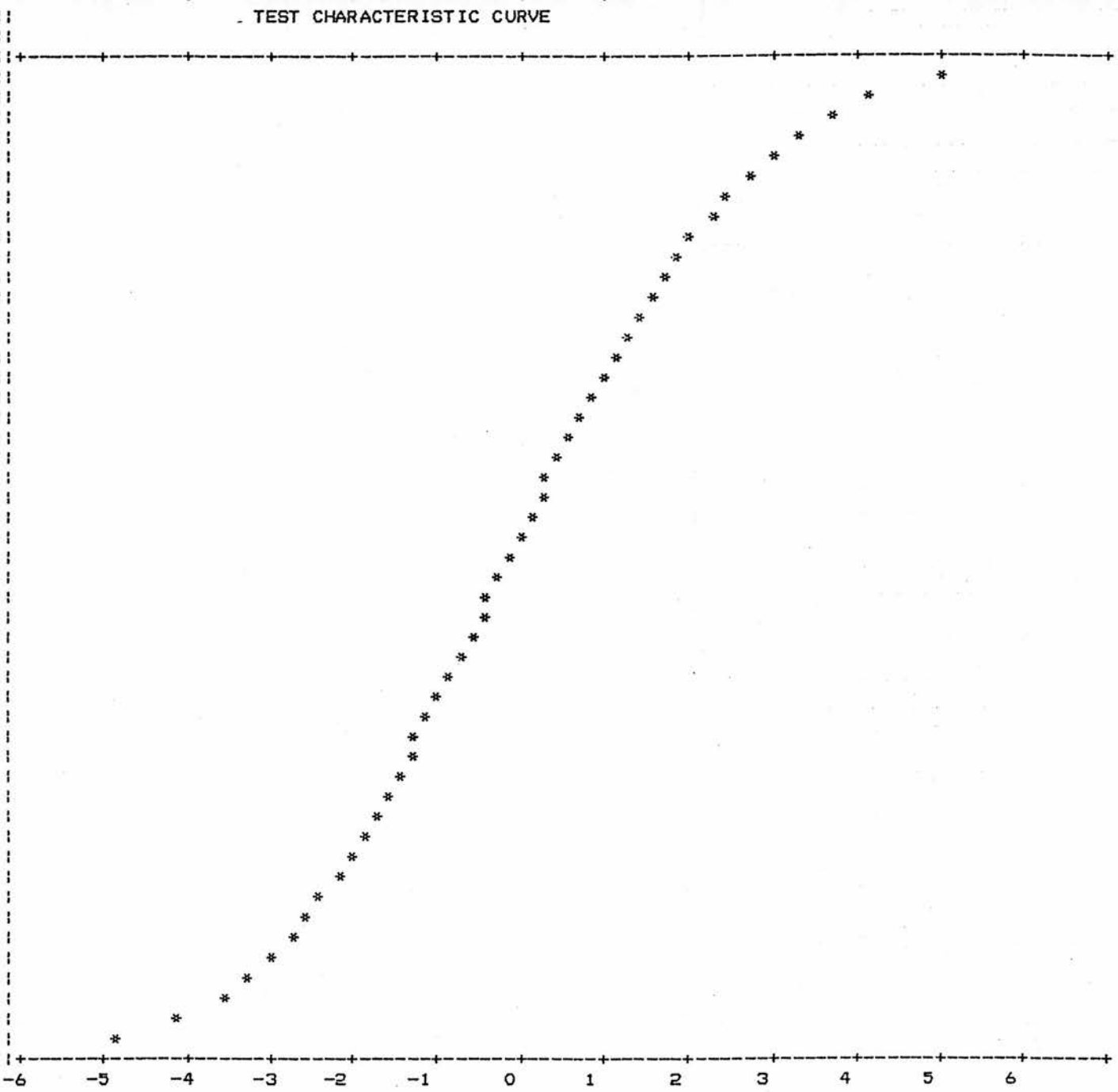
7.5. Ability: using Rasch analysis with people

7.5.1. Relationship between difficulty and ability

One of the features of the Rasch model, as with other IRT models, is that difficulty and ability are reported on the same scale. A full exploration of this issue would be needed in any interesting implementation of the item banking concept but will not be developed here. However, the table reported here for Form A shows how ability and difficulty are related to each other, and where the raw score fits in (full analyses of the other Forms are available from the author).

What should be noted is the way in which a candidate can be matched with items which are suitable for him: thus a candidate with ability −0.50, say, (corresponding to a raw score on the total test of 57) would be 'best tested' by items 12, 36 or 71. In an adaptive testing procedure, that candidate could be routed to items just above or below this level as his ability estimate is revised. This technical development will not be pursued here, though it should be noted how adaptive testing, tailored testing etc. grow out of the fundamentals of item banking.

RAW SCORE	COUNT	LOG ABILITY	STANDARD ERRORS
49	0	5.04	1.06
48	1	4.25	0.79
47	1	3.75	0.66
46	4	3.37	0.59
45	1	3.05	0.54
44	6	2.79	0.51
43	10	2.55	0.48
42	8	2.34	0.46
41	7	2.14	0.44
40	10	1.96	0.42
39	15	1.79	0.41
38	18	1.63	0.40
37	12	1.48	0.39
36	12	1.34	0.38
35	8	1.20	0.38
34	12	1.06	0.37
33	18	0.93	0.37
32	14	0.80	0.36
31	8	0.67	0.36
30	11	0.55	0.36
29	11	0.42	0.35
28	10	0.30	0.35
27	6	0.18	0.35
26	11	0.06	0.35
25	10	-0.06	0.35
24	3	-0.18	0.35
23	6	-0.30	0.35
22	1	-0.41	0.35
21	3	-0.53	0.35
20	2	-0.65	0.35
19	1	-0.77	0.35
18	4	-0.90	0.35
17	1	-1.02	0.36
16	0	-1.15	0.36
15	1	-1.27	0.36
14	1	-1.40	0.37
13	1	-1.54	0.37
12	0	-1.68	0.38
11	0	-1.82	0.39
10	0	-1.98	0.40
9	0	-2.14	0.41
8	0	-2.31	0.43
7	0	-2.50	0.45
6	1	-2.71	0.47
5	0	-2.94	0.51
4	0	-3.22	0.56
3	0	-3.56	0.63
2	0	-4.01	0.75
1	0	-4.75	1.04



PERSON SEPARABILITY INDEX 0.85 (EQUIVALENT TO KR20)

50 ITEMS CALIBRATED ON 248 PERSONS

Table 14. Relationship between difficulty and ability (Form A)

7.5.2. Deriving and comparing ability estimates

In the same way that difficulty estimates were obtained both from the whole test and from the individual parts (see section 7.3), so ability estimates can be obtained in the same way. Actual ability estimates are not given here but are obtainable from the author. Although ability is clearly very important in any use to which an item bank is to be put - perhaps the most important element, since it is the purpose of the testing procedure to obtain estimates of ability - we are concerned here with the design and construction of an item bank and therefore, in so far as this is possible, attempt to focus on difficulty estimates rather than ability estimates. This means that actual ability estimates, while being used in the analysis and in the construction of the bank, will not be of direct relevance in themselves.

Ability estimates for all testees were obtained from the whole test and from the separate parts. These estimates were then correlated to see how consistent ability estimates would be, especially in view of the fact that Part 3 of the test on all forms was considerably more difficult than the other parts. The correlation coefficients obtained were as follows:

Table 15
Correlation of ability estimates derived from sub-tests and whole test

Form A

	Part 1	Part 2	Part 3	Whole test
Part 1	1.00	.78	.67	.93
Part 2	-	1.00	.64	.90
Part 3	-	-	1.00	.85

Form B

Part 1	1.00	.81	.75	.95
Part 2	-	1.00	.71	.92
Part3	-	-	1.00	.87

Form C

Part 1	1.00	.82	.98	.96
Part 2	-	1.00	.81	.92
Part 3	-	-	1.00	.94

Form D

Part 1	1.00	.77	.79	.95

Part2	-	1.00	.72	.92
Part 3	-	-	1.00	.88

The part-whole correlations are high enough for us to conclude that abilities estimated on parts of the test will be linearly related to abilites estimated on the basis of the whole test - much in the way that difficulties so estimated are related. There is a warning lurking in these figures, however, and that is that in some cases estimates obtained from one part of the test will not be closely related to estimates obtained from another part of the test. This suggests that distortions in estimates can occur, in this case if Part 3 estimates are taken alone. The reason for this is that on the whole Part 3 was the most difficult part of the test and many testees will have been responding to items at a level too high for their ability. This confirms the desirability, made possible by a properly calibrated bank of items, of matching a testee's ability to the difficulty of the item he takes.

7.5.3. An item bank: 1

It is clear that any one of the test forms taken in isolation could form the basis for an item bank. In other words any one test form could be an item bank, since it consists of a fully calibrated set of items on a unidimensional common scale. In this sense we have, then, already constructed an item bank, though not a very large or particularly interesting one, simply by the act of producing a table such as that given in section 7.4. By taking the four different test forms we would be able to present four different item banks.

Now, while this is a satisfactory first step, it fails to make use of all the items we have available, and thus fails to combine items from different test forms on a common scale. We do not propose, therefore, to dwell on these isolated banks.

However, as a matter of principle it should be noted that even at this crude first stage there still remains the task of grouping together those items which are interdependent, namely the Part 2 items identified as INT 10 - 13 and the Part 3 comprehension passages. This task will be discussed in the next section.

7.6. Developing an item bank

The next stage is to put together the information on the items that we have been analysing so that the items can be calibrated on a common scale. The methods for doing this have already been outlined; what follows is a practical example of how

such methods would be used on the kind of data we have been looking at. The principle is simple: given that the items have been shown to exhibit a high degree of unidimensionality, we may now proceed to equate the tests (either Form A with Form B or Form C with Form D) so as to obtain a single bank of 280 items calibrated on a common scale. In order to show the principles involved, the next section demonstrates how the procedure works, taking as its 'anchor test' (see chapter 5) a sub-set of the Form A items which were used in the pre-pilot phase while the test was being constructed.

7.6.1. Calibration using a high ability group

The items which were used in the pre-pilot phase and which were subsequently used in the full pilot phase were calibrated on a Rasch scale and treated as the 'anchor test' for calibrating the rest of the items on Form A. The calibration of these 'anchor' items was done manually using the PROX procedure. Since the pre-pilot group was a group of teachers at USM we can consider them to be a 'high ability' group. The result of the recalibration procedure is shown in the following table:

Table 16
Calibration procedure using a high ability group as the anchor

Item No.	FV	d(PROX)	d(UCON)	Difference	Combined (a)	Combined (b)	Adjusted d
A70	94	-1.50	+1.537	+3.037	+1.294	-O.103	+0.188
A62	94	-1.50	-1.758	-0.258	-2.001	-1.751	-1.460
A73	89	-0.57	+0.761	+1.331	+0.258	-0.156	+0.135
A51	56	+3.67	+0.694	-2.976	+0.451	+2.061	+2.352
A71	83	+0.13	+0.130	-0.000	-0.113	+0.074	+0.365
A72	89	-0.57	-1.320	-0.750	-1.563	-1.067	-0.776
A67	94	-1.50	-0.368	+1.132	-0.611	-1.056	-0.765
A54	89	-0.57	+0.270	+0.840	+0.027	-0.272	-0.019
A68	94	-1.50	-1.117	+0.383	-1.360	-1.430	-1.139
A61	94	-1.50	-1.081	+0.419	-1.324	-1.412	-1.121
A65	56	+3.67	+2.978	-0.692	+2.735	+3.023	+3.314
A55	61	+1.72	+1.404	-0.316	+1.161	+1.441	+1.732
A57	78	+0.57	+1.538	+0.968	+1.295	+0.933	+1.224
A59	28	+4.58	+3.758	-0.822	+3.515	+4.048	+4.339
A63	83	+0.13	+1.767	+1.637	+1.524	+0.187	+0.478
A74	94	-1.50	-0.069	+1.431	-0.312	-0.906	-0.615
A75	89	-0.75	+0.345	+1.095	+0.102	-0.324	-0.033
A78	94	-1.50	-0.298	+1.202	-0.541	-1.021	-0.730
A79	89	-0.57	+0.294	+0.864	-0.537	-0.554	-0.263
A80	83	+0.13	+0.319	+0.189	+0.076	+0.103	+0.394
A81	83	+0.13	+0.394	+0.264	+0.151	+0.141	+0.432
A95	83	+0.13	+0.320	+0.190	+0.077	+0.104	+0.395
A97	94	-1.50	+1.442	+0.058	+1.199	-0.151	+0.140
A99	17	+4.58	+2.431	-2.149	+2.188	+3.384	+3.675

A85	61	+2.02	−0.941	−2.961	−1.184	+0.418	+0.709
A86	89	−0.57	−0.740	−0.170	−0.983	−0.538	−0.247
A87	72	+1.04	−0.262	−1.302	−0.505	+0.268	+0.559
A88	89	−0.57	−1.476	−0.906	−1.719	−1.145	−0.854
A90	89	−0.57	+1.734	+2.304	+1.491	+0.461	+0.752
A91	94	−1.50	+1.734	+3.234	+1.491	−0.005	−0.286
A52			−1.247		−1.490		−1.199
A53			−2.082		−2.325		−2.034
A56			−2.134		−2.377		−2.086
A58			−2.294		−2.537		−2.246
A60			+1.209		+0.966		+1.257
A64			−0.788		−1.031		−0.740
A66			−0.740		−0.983		−0.692
A69			+1.231		+0.988		+1.279
A76			+4.068		+3.825		+4.116
A77			+3.431		+3.188		+3.479
A82			−0.889		−1.133		−0.842
A83			−0.720		−0.963		−0.672
A84			−0.269		−0.512		−0.221
A89			−1.283		−1.526		−1.235
A92			−1.656		−1.899		−1.608
A93			−2.293		−2.536		−2.245
A94			−3.413		−3.656		−3.365
A96			−1.934		−2.177		−1.886
A98			−2.134		−2.377		−2.086
A100			−0.483		−0.726		−0.435
Mean		0.01	0.00	0.243	−0.260	−0.291	0.027

The Pearson product moment correlation coefficient for FV and d(PROX) is 0.92, while the correlation coefficient for d(PROX) and d(UCON) is 0.57.

The average difficulty of the 30 link items increased by 0.243; this amount is thus subtracted from all values obtained in the pilot (longer) set of items (since we want a figure between the anchor value and the pilot value). The combined (a) value is averaged with the the original anchor value to give the combined (b) value. The adjusted value is the final combined value +0.291 to centre values on zero (though there is a slight rounding error which gives rise to a mean of slightly greater than zero).

The effect of this procedure is simply to adjust difficulty estimates up or down by the addition or subtraction of a constant. It can easily be seen that if enough calibrations and recalibrations were performed, then the final estimate of the difficulty parameter for each item would tend towards the mean value of that estimate for all the occasions on which it was obtained. The rather extreme example given above contains several quite dramatic changes in estimate (item A70, for example, has to be calibrated from the extreme values of −1.50 and +1.54), but it can still be seen how rapidly estimates will tend to gather round a mean upon recalibration. In practice, as

will be seen, calibration is more a process of fine tuning than the reconciliation of opposites.

What we require above all else in calibration procedures of this kind is confidence that items have been compared consistently and in accordance with our views of the underlying construct. It is interesting to note that in the calibration described in this section, the extreme difficulty values arise precisely in those items which we have already identified as being 'odd' in some way - whether because they were poorly written or because they appear to depend on some sort of learning effect having taken place. Teachers, for example, may be very capable of teaching 'text recognition' and 'information transfer' in the sense in which these terms are used here, but they perform quite poorly in relation to their pupils on tests of these items. This raises a number of interesting questions which could be explored elsewhere.

7.6.2. Comparison of obtained calibrated values

The method shown above was used to calibrate the items in Forms C and D, first using the common items from Form A as the anchor set and then using the common items from Form B as the anchor set. Thus three values were obtained for the 'difficulty' of each item. The results for each part of the test are shown separately below. In this analysis, each part was calibrated as if it were a separate test - i.e. the assumption is made that a separate scale is needed for each part of the test.

Table 17
Summary of obtained difficulty values for Forms C and D: Part 1

Item Label	With Form A items as anchor set	With Form B items as anchor set	Observed value
C1	0.34	0.63	0.31
2	-1.29	-0.77	-1.04
3	0.49	0.54	0.52
4	-0.37	-0.53	-0.43
5	-1.16	-0.23	-0.81
6	0.30	-0.06	0.14
7	-1.52	-0.59	-1.45
8	1.42	1.44	1.24
9	0.84	1.21	1.06
10	0.53	0.62	0.86
11	-1.99	-1.26	-2.13
12	1.13	0.94	1.08
13	-0.17	0.58	0.46
14	0.61	0.54	0.96
15	1.58	0.97	1.08

16	−0.71	−0.16	−0.02
17	−1.29	−0.56	−0.96
18	−1.74	−1.15	−1.90
19	−0.42	0.50	0.14
20	1.83	2.44	1.85
21	−2.23	−2.43	−2.03
22	2.04	1.19	1.58
23	−1.47	−1.76	−1.55
24	−1.80	−1.18	−1.94
25	1.36	2.00	1.30
26	−0.59	−0.45	−0.89
27	3.29	3.53	3.18
28	−1.42	−0.39	−0.79
29	−1.47	−0.68	−1.04
30	0.32	1.30	0.71
31	3.81	1.10	1.92
32	−0.68	−0.74	−0.45
33	−1.99	0.42	0.16
34	0.90	0.87	0.52
35	−0.87	−1.00	−1.12
36	−1.92	−1.66	−1.15
37	−3.26	−3.10	−3.05
38	3.60	−0.45	−0.16
39	1.54	2.21	1.58
40	1.60	1.82	1.30
41	0.34	0.16	0.62
42	−0.45	0.02	−0.51
43	−1.74	−1.04	−0.84
44	1.50	1.33	1.26
45	0.59	0.42	0.43
46	2.27	0.72	1.41
47	−1.74	−0.39	−0.86
48	−1.74	−1.04	−1.42
49	1.60	1.08	1.06
50	0.17	−0.18	−0.20
D1	−1.30	−1.43	−1.38
2	−0.89	−0.74	−0.77
3	1.95	1.95	1.92
4	0.18	0.08	0.13
5	0.78	0.61	0.58
6	0.43	2.44	2.41
7	−1.26	−1.18	−1.39
8	−0.85	−0.98	−1.21
9	−3.15	−1.87	−1.90
10	−0.26	−0.39	−0.34
11	2.00	1.87	1.92
12	1.01	0.83	0.80
13	−0.29	−0.42	−0.37
14	2.25	2.12	2.17
15	0.03	−1.00	−1.03
16	−0.91	−1.00	−1.03
17	−0.58	−0.71	−0.66
18	−1.06	−1.19	−1.14
19	−1.80	−1.75	−1.78
20	−0.66	−0.79	−0.74
21	1.36	1.23	1.28

22	−2.04	−2.17	−2.12
23	0.42	0.22	0.19
24	−0.69	0.12	0.17
25	−0.69	−0.82	−0.77
26	0.80	0.58	0.55
27	−0.62	−0.15	−0.18
28	−0.63	−0.76	−0.71
29	−0.08	−0.24	−0.27
30	0.45	0.32	0.37
31	−1.26	−0.77	−0.80
32	0.61	0.48	0.53
33	−1.39	−1.52	−1.47
34	−1.13	−1.26	−1.21
35	1.30	1.17	1.22
36	0.95	0.82	0.87
37	0.14	0.01	0.06
38	3.76	3.34	3.31
39	−0.20	−0.08	−0.11
40	−0.82	−0.49	−0.52
41	0.75	0.77	0.74
42	0.29	0.16	0.21
43	−0.32	−0.52	−0.55
44	1.90	2.02	1.99
45	0.53	0.40	0.45
46	0.80	0.77	0.74
47	−0.05	−0.19	−0.22
48	−1.39	−1.52	−1.47
49	−0.25	−0.15	−0.18
50	1.07	1.53	1.50

Table 18
Summary of obtained difficulty values for Forms C and D: Part 2

Item Label	With Form A items as anchor set	With Form B items as anchor set	Observed value
C51	1.32	1.69	1.66
52	−1.77	−0.98	−1.19
53	0.55	0.66	0.67
54	−0.58	−0.60	−0.45
55	1.41	1.20	1.07
56	−0.09	0.49	0.29
57	1.25	1.76	1.83
58	0.73	0.25	0.78
59	1.74	1.63	1.33
60	0.35	0.20	0.35
61	0.29	−0.01	0.39
62	0.85	0.06	0.91
63	0.07	0.03	0.11

64	−1.41	−1.08	−1.15
65	0.29	0.10	0.37
66	−1.69	−0.92	−1.11
67	−0.48	−0.38	−0.56
68	−0.89	−0.81	−0.86
69	4.36	4.65	5.15
70	−1.35	−1.52	−1.74
71	−0.04	−0.76	−0.18
72	−0.38	0.48	0.82
73	2.91	1.76	1.66
74	0.33	1.68	1.64
75	1.59	0.55	0.82
76	1.21	0.18	0.87
77	3.81	3.04	3.27
78	−0.35	−0.11	−0.15
79	0.44	0.18	0.39
80	−0.14	0.29	0.50
81	−0.73	−0.25	0.62
82	−0.77	−0.89	−1.32
83	0.35	0.27	0.56
84	0.62	1.41	0.74
85	−0.98	−1.04	−1.11
86	−0.62	−0.41	−0.62
87	−0.89	−0.58	−0.45
88	−1.96	−1.07	−2.08
89	−2.17	−2.84	−3.00
90	−2.59	−2.10	−2.57
91	−2.43	−1.20	−2.16
92	−2.29	−2.25	−3.00
93	−2.59	−2.70	−3.19
94	−3.48	−3.33	−3.42
95	1.62	1.10	1.08
96	1.43	1.21	1.28
97	1.36	0.82	1.30
98	−0.41	−0.03	−0.22
99	0.59	0.99	0.78
100	−1.29	−0.58	−0.70

D51	−1.69	−1.62	−1.70
52	−0.83	−0.74	−0.84
53	1.19	1.28	1.18
54	−0.05	0.04	−0.06
55	2.56	2.65	2.55
56	1.10	1.19	1.09
57	0.47	0.56	0.46
58	0.64	0.73	0.63
59	−1.07	−1.02	−1.00
60	0.92	0.97	0.99
61	−0.94	−0.91	−0.87
62	1.36	1.45	1.35
63	−1.77	−1.72	−1.70
64	2.85	2.90	2.92
65	−1.04	−0.99	−0.97
66	0.06	0.11	0.13
67	0.74	0.79	0.81
68	−1.38	−1.33	−1.31

69	0.64	0.69	0.71
70	−1.01	0.96	−0.94
71	−0.29	−0.20	−0.30
72	−0.91	−0.82	−0.92
73	3.28	3.33	3.35
74	−0.05	0.00	0.02
75	−0.40	−0.31	−0.41
76	1.78	1.85	1.85
77	0.78	−0.43	0.77
78	−0.53	−0.44	−0.54
79	−0.69	−0.60	−0.70
80	−0.36	−0.27	−0.37
81	−1.07	−0.98	−1.08
82	−1.23	−1.18	−1.16
83	−0.33	−0.28	−0.26
84	−0.63	−0.58	−0.56
85	0.55	0.66	0.56
86	−0.15	−0.06	−0.16
87	−0.09	0.00	−0.10
88	−0.09	0.00	−0.10
89	0.98	1.07	0.97
90	0.01	0.10	0.00
91	−0.33	−0.24	−0.34
92	−0.36	−0.27	−0.37
93	1.34	1.43	1.33
94	1.04	1.13	1.03
95	−0.35	−0.30	−0.28
96	−1.58	−1.53	−1.51
97	0.74	0.35	0.81
98	−3.53	−3.48	−3.46
99	1.07	1.12	1.14
100	−1.21	−1.43	−1.44

Table 19
Summary of obtained difficulty values for Forms C and D: Part 3

Item Label	With Form A items as anchor set	With Form B items as anchor set	Observed value
C101	−1.63	−1.50	−1.24
102	−2.15	−2.34	−2.02
103	0.89	0.05	0.11
104	−2.39	−3.12	−2.49
105	0.49	0.31	0.31
106	0.95	0.26	0.40
107	−0.34	−0.42	0.12
108	0.67	0.86	0.58
109	−0.45	−0.35	−0.77
110	1.22	1.44	0.75
111	1.80	1.06	1.01
112	−0.17	−1.54	−1.51

113	0.59	0.28	0.19
114	−0.34	−0.68	−0.28
115	0.49	0.89	0.94
116	0.40	0.32	0.31
117	1.80	2.12	2.18
118	0.33	0.28	0.23
119	0.48	0.34	0.51
120	0.55	0.13	0.49
121	−1.07	−0.60	−0.83
122	0.19	0.61	0.96
123	−0.38	−0.53	−0.74
124	0.59	1.01	0.81
125	−0.50	−0.60	−0.10
126	−1.22	−1.13	−1.35
127	−0.50	−0.39	−0.42
128	−0.87	−0.70	−0.77
129	1.22	1.10	1.03
130	−1.77	−1.96	−1.44
131	0.57	0.53	0.23
132	0.40	0.32	0.92
133	−0.34	−0.31	−0.94
134	0.55	0.46	0.60
135	1.56	1.66	1.75
136	0.57	0.24	0.62
137	1.97	1.70	1.59
138	−0.87	−0.70	−0.37
139	−0.54	−0.08	−0.42
140	−0.95	−1.04	−0.92
D101	−0.68	−0.15	−0.96
102	0.53	0.50	0.01
103	−0.08	−0.25	−0.02
104	−1.04	−0.91	0.16
105	0.32	0.82	0.37
106	−0.39	−0.40	−1.17
107	−0.13	−0.33	−0.01
108	1.12	1.69	−0.58
109	−0.60	−0.91	−0.17
110	−0.35	0.06	0.44
111	0.28	0.16	−0.22
112	0.05	0.18	−0.89
113	0.46	0.52	0.76
114	−1.96	−1.03	−0.36
115	0.24	0.10	−0.29
116	−0.70	−0.58	1.42
117	0.26	0.22	0.29
118	−0.60	−0.99	−0.82
119	−3.12	−2.15	−2.23
120	−0.66	−1.55	−1.28
121	−0.31	−0.60	−0.55

122	-0.10	-0.58	-0.49
123	-0.70	-0.10	-0.06
124	0.75	0.20	0.27
125	0.53	1.00	0.97
126	-0.42	-0.31	-0.20
127	-0.31	-0.87	-0.84
128	-1.50	-1.15	-0.98
129	1.35	0.35	0.40
130	1.06	1.47	1.45
131	0.28	-1.11	-1.05
132	0.89	-1.90	-1.50
133	0.61	0.96	0.90
134	1.01	1.45	1.18
135	-0.34	0.37	0.33
136	0.86	1.45	1.39
137	0.44	1.42	1.30
138	0.13	0.54	0.76
139	1.54	1.32	1.21
140	0.32	1.35	1.08

The correlation between difficulty estimates obtained by using Form A as the anchor test and those obtained by using Form B is 0.93. We can therefore be confident that both sets of anchor items are working in harmony.

Inspection of the above table shows how the calibration process essentially works towards the production of an average value for estimates of item parameters. The operation is basically an arithmetic one of no controversy, provided that the underlying assumptions of the model used are met.

7.6.3. An item bank: 2

We are now in a position to order the items into a single bank. For convenience we take Form A as the anchor test, though as demonstrated in the previous section it would be just as acceptable to use Form B as the anchor test. The most important procedure now remaining is to group together those items which are inter-dependent. We could do this by identifying those groups (INT 10 - 13 and COMP 1 - 7) and taking the average difficulty value for the items in those groups. This average difficulty would then be used as the item bank difficulty estimate. The values of the other items in the bank would need to be adjusted slightly (by a constant amount) to take into account the loss of test information generated by this grouping procedure.

However, in the calibration given here, and in the item bank presented in this

chapter, the individual item difficulties are retained because it is possible that a user of the bank may wish to use a comprehension passage, say, with just one item e.g. to test 'understanding the main idea'. This is a justifiable, though time-consuming and perhaps wasteful, use of items in a bank.

So although one constraint upon the use of items in a bank is that they be independent of each other, for the purposes of the present study such a grouping was not carried out in order that the full information for each individual item be available. An averaged method, as just outlined, or a partial credit method (Wright and Masters 1982) would result in loss of information which could obscure the detailed picture we have built at this point. For most practical purposes, however, such a grouping is advisable.

For the final estimates of the item bank difficulty figures the initial estimates from the whole test analysis were used. This differs from the part test figures used in the previous section, mainly for the reason that we are assuming unidimensionality and should therefore use the more accurate whole test figures (though as demonstrated in section 7.4 there is a high correlation between these figures). Another important reason for using the whole test analysis at this stage is that once we proceed to group items together, then the loss of information in any one part of the test, particularly Part 3, may distort the values obtained.

Finally, some way of reporting the results of using the bank needs to be given. If the objective is simply to discover a testee's ability for placement purposes, then no further action is necessary. As we have seen in section 7.4, the difficulty scale corresponds exactly to the ability scale.

However, if results need to be reported to testees or to outsiders then it is necessary to go a step further. The reason for this is that it makes little sense to tell someone that they have an ability of, say, −1.52. The simplest way of producing an interpretable *number* is simply to impose an arbitrary scale alongside the ability scale. This has the advantage of being easy to perform and of being adaptable to the test user's requirements: a scale of 100 is just as easy to impose as a scale of 500, or whatever. It has the disadvantage of being arbitrary and of concealing the real meaning of the ability estimate. It is also of no use in directing future test activity, since it would have to be reconverted to the ability scale in order to be comparable with the difficulty scale.

This is not an issue which is easily resolvable, since it depends on considerations of practical administration and of public perception of what numbers derived from

tests actually mean. As an interim solution, the use of a consumer-specified scale seems as good as any. Choppin's (1979) solution is to create an ability scale which is itself interpretable, in the sense that it corresponds to a range with which consumers feel comfortable, in his case between 1 and 10. Haksar (1984) adopts a similar approach. However, the problem here is that consumers are still using scales in ways which are meaningful to them, and the attempt by test constructors to change a basically unhelpful scale into a helpful one merely by manipulating the underlying figures tends to obscure the fact that the test constructor is still using the test in one way while the consumer is using it (or at least interpreting the figures derived from it) in another way. This is a debate which we do not attempt to resolve here.

Another step to be taken is to say what these ability figures actually mean. We are after all claiming that ability is in some sense a criterion-referenced concept, and that therefore we should be able to say what it means when we say that a testee has an ability of such-and-such. Enough has been said already to conclude that in fact we can say little more than that a testee has a certain amount of ability in a general construct called 'reading in English as a foreign language'. Our attempts to identify separate strands and elements of this generalised ability produced no positive results. However, it is reasonable to continue to label items in accordance with current thinking in the field and to say that certain items test 'grammar', others test 'reading comprehension' and so on; furthermore, we could label item types as 'understanding of prepositions', say, or 'recognising sequence' and so on. This would be a particularly useful procedure to adopt in any situation where retrieval of items by content description (see chapter 4) was felt to be necessary. The danger is that such labelling tends to suggest categories of the construct which are not really identifiable. Again, the decision to label must remain a consumer's decision. The present author feels that this should not be done with the field under investigation, and prefers to retain the broad category of 'reading EFL'.

A 'map' of the item bank is given in the next two tables, which present the items in Form C and Form D separately for convenience, though they are calibrated on a common scale.

7.7. Conclusion

We have seen that the Rasch model fits data of the type analysed here and that such a model can therefore be used with reasonable confidence to proceed with practical applications.

On the basis of the results given here it would appear that there is no justification

PERSON STATS	COUNT	RAW SCORE	MEASURE MIDPOINT (S.E.)	ITEM COUNTS	TYPICAL ITEMS (BY NAME)									
			4.50 (0.62)	1	C 69									
		136	4.30 (0.54)											
+4SD		135	4.10 (0.49)											
		134	3.90 (0.45)											
		133	3.70 (0.42)											
		132	3.50 (0.39)											
+3SD		130	3.30 (0.36)											
	1	128	3.10 (0.33)	1	C 27									
		126	2.90 (0.31)	2	C 77	C117								
	4	124	2.70 (0.29)											
	3	121	2.50 (0.27)	1	C135									
+2SD	3	118	2.30 (0.26)	1	C137									
	8	115	2.10 (0.25)											
	11	111	1.90 (0.23)	1	C 31									
	10	108	1.70 (0.23)	7	C 20	C 39	C111	C115	C122	C129	C132			
+1SD	21	103	1.50 (0.22)	4	C 22	C 46	C110	C124						
	16	99	1.30 (0.21)	11	C 8	C 25	C 40	C 44	C 57	C 73	C108	C119	C120	C134
	21	94	1.10 (0.21)	9	C 9	C 12	C 15	C 49	C 51	C 74	C105	C106	C116	
	15	89	0.90 (0.20)	9	C 10	C 14	C 30	C 59	C 97	C103	C113	C118	C131	
MEAN	17	84	0.70 (0.20)	3	C 41	C 96	C107							
	9	79	0.50 (0.20)	9	C 3	C 13	C 34	C 45	C 55	C 62	C 76	C 95	C125	
	24	74	0.30 (0.20)	13	C 1	C 6	C 19	C 33	C 58	C 72	C 75	C 84	C 99	C114
	26	69	0.10 (0.20)	4	C 16	C 53	C 81	C 83						
-1SD	21	64	-0.10 (0.20)	9	C 38	C 50	C 65	C 79	C 80	C109	C121	C123	C128	
	12	59	-0.30 (0.20)	6	C 4	C 56	C 60	C 61	C133	C140				
	18	54	-0.50 (0.20)	3	C 32	C 42	C 63							
	8	49	-0.70 (0.21)	10	C 5	C 26	C 28	C 43	C 71	C 78	C 98	C101	C112	C126
	2	45	-0.90 (0.21)	4	C 17	C 29	C 47	C130						
-2SD		41	-1.10 (0.22)	6	C 2	C 35	C 36	C 54	C 67	C 87				
	1	37	-1.30 (0.23)	2	C 86	C100								
		33	-1.50 (0.23)	5	C 7	C 23	C 48	C 68	C 85					
		29	-1.70 (0.25)	4	C 18	C 24	C 52	C102						
-3SD		26	-1.90 (0.26)	4	C 11	C 64	C 66	C 82						
		23	-2.10 (0.27)	2	C 21	C104								
		21	-2.30 (0.28)	1	C 70									
		18	-2.50 (0.29)											
-4SD		16	-2.70 (0.31)	1	C 88									
		14	-2.90 (0.33)	1	C 91									
		12	-3.10 (0.35)	1	C 37									
		11	-3.30 (0.36)	2	C 90	C 92								
-5SD		9	-3.50 (0.39)											
		8	-3.70 (0.41)	1	C 93									
		7	-3.90 (0.44)											
		6	-4.10 (0.47)	1	C 89									
		5	-4.30 (0.51)											
			-4.50 (0.51)											
		4	-4.70 (0.56)											
		3	-4.90 (0.64)											
			-5.10 (0.64)	1	C 94									
			-5.30 (0.64)											

140 ITEMS CALIBRATED ON 251 PERSONS
251 MEASURABLE PERSONS WITH MEAN ABILITY = 0.69 AND STD. DEV. = 0.85

FORM C WHOLE TEST

Map of Items for the Item Bank

PERSON STATS	COUNT	RAW SCORE	MEASURE MIDPOINT (S.E.)	ITEM COUNTS	TYPICAL ITEMS (BY NAME)									
+4SD		131	3.30 (0.36)	1	D 73									
		129	3.10 (0.34)											
		127	2.90 (0.31)	1	D 64									
		125	2.70 (0.30)	1	D 38									
+3SD	1	122	2.50 (0.28)	1	D 55									
	3	120	2.30 (0.27)	2	D116	D130								
	1	116	2.10 (0.25)	3	D134	D136	D137							
+2SD	5	113	1.90 (0.24)	3	D 6	D139	D140							
	7	109	1.70 (0.23)	3	D 14	D 76	D125							
	13	105	1.50 (0.22)	4	D 44	D113	D133	D138						
	3	101	1.30 (0.21)	4	D 3	D 11	D 62	D 93						
+1SD	14	96	1.10 (0.21)	8	D 53	D 56	D 99	D105	D110	D117	D129	D135		
	12	92	0.90 (0.20)	6	D 50	D 60	D 89	D 94	D104	D124				
	21	87	0.70 (0.20)	10	D 21	D 35	D 67	D 69	D 77	D 97	D102	D103	D107	D123
	22	82	0.50 (0.20)	5	D 58	D 85	D109	D115	D126					
MEAN	28	76	0.30 (0.19)	7	D 12	D 36	D 41	D 46	D 57	D111	D114			
	22	71	0.10 (0.19)	7	D 5	D 26	D 32	D 66	D108	D121	D122			
	19	66	-0.10 (0.19)	8	D 30	D 45	D 54	D 74	D 87	D 88	D 90	D127		
	26	60	-0.30 (0.20)	14	D 4	D 23	D 24	D 42	D 71	D 83	D 86	D 95	D101	D106
-1SD	19	55	-0.50 (0.20)	5	D 37	D 75	D 80	D 91	D 92					
	6	50	-0.70 (0.20)	9	D 27	D 29	D 39	D 47	D 49	D 78	D 79	D 84	D120	
	8	45	-0.90 (0.20)	5	D 10	D 13	D 52	D 61	D 72					
-2SD	4	40	-1.10 (0.21)	8	D 17	D 40	D 43	D 59	D 65	D 70	D 81	D132		
	2	36	-1.30 (0.22)	7	D 2	D 20	D 25	D 28	D 31	D 68	D 82			
		32	-1.50 (0.22)	5	D 8	D 15	D 16	D 96	D100					
		28	-1.70 (0.23)	6	D 7	D 18	D 34	D 51	D 63	D119				
-3SD		24	-1.90 (0.25)	3	D 1	D 33	D 48							
		21	-2.10 (0.26)											
		18	-2.30 (0.27)	1	D 19									
		16	-2.50 (0.29)	1	D 9									
-4SD		14	-2.70 (0.30)	1	D 22									
		12	-2.90 (0.32)											
		10	-3.10 (0.35)											
-5SD		8	-3.30 (0.38)											
		7	-3.50 (0.40)	1	D 98									
		6	-3.70 (0.43)											

140 ITEMS CALIBRATED ON 236 PERSONS
236 MEASURABLE PERSONS WITH MEAN ABILITY = 0.34 AND STD. DEV. = 0.74

FORM D WHOLE TEST

for thinking of reading in English as a foreign language as anything other than unidimensional. No evidence was found for the existence of 'subskills', nor was there strong evidence to suggest that an all-embracing definition of 'reading in English as a foreign language' which included atomistic grammar-recognition tasks as well as more global reading comprehension tasks would be untenable.

Items of the type used in this study can be calibrated using different anchor sets and similar results will be achieved. An item bank is thus a workable concept, whatever use the test developer wishes to make of it.

We now look at the shortcomings of the present investigation and go on to suggest future developments and the potential for item banking in foreign language learning.

CHAPTER 8

CONCLUSION: ITEM BANKS AND EFL READING

8.1. Introduction

This final chapter recapitulates the main points made earlier and suggests the direction in which future research may go.

8.2. Limitations of the present study

Several comments need to be made about the shortcomings of the present study. Firstly, the practical work described here is limited to one group of learners in a particular environment (Malaysian undergraduates at the matriculation phase of their education). If the concept of 'sample-freeness' is to be validated more rigorously, it will be necessary to use the items piloted here on other more diverse groups. To a certain extent this problem has been explored with already existing test data (Woods and Baker, 1986; Baker 1987) and it has been found that Rasch estimates of difficulty are relatively stable across different populations. The problem has not been fully explored, however, in that the tests used in such analyses were not constructed according to strict, publicly available criteria, such as has been attempted here.

It may be argued that Malysian undergraduates are not, linguistically, a homogeneous group: there are after all speakers of Malay as well as Chinese and Indian languages as a mother tongue represented in the undergraduate population. To that extent, the data may be said to be derived from a heterogeneous population. However, the method of analysis has been to treat the group homogeneously and not to subdivide according to mother tongue (which may be a shortcoming of the experimental design). It therefore still remains to be seen whether across different linguistic groups 'sample-freeness' still holds for strictly constructed tests. Perhaps it is the culture rather than the language of the population which will prove to be the important factor. The argument tends to move towards a discussion of language proficiency and whether there can be such a thing as an 'absolute' proficiency in language, or any aspect of it, independent of the culture within which it is being studied. This is also partly a problem of educational definitions: if 'reading for the main idea' is deemed to be part of the reading proficiency of a student in one educational context but not in another, then it is hard to see how a test item of this type will remain sample free.

A second limitation of the present study is that it is limited to the design and construction stages of item banking. No attempt has been made to put the bank into

practice at this stage. Having shown that it is possible and defensible to construct a bank, it now needs to be show how this would work in practice – in adaptive testing, for example. This is not an issue which is going to be solved purely on theoretical grounds; the practice of item banking is the criterion by which it should be judged.

In connection with the construction of the present bank it should also be noted that it is largely the work of one person and is therefore subject to the inevitable limitations that such a procedure represents.

Thirdly, there are a number of considerations in the design of the tests used to develop the current bank which may have a limiting influence on the conclusions we are able to draw. The balance of the parts of the test may be wrong, for instance. It is quite possible that the large number of single sentence multiple choice questions tended to produce too easy a test. While this should not really matter, in so far as it was intended that a wide range of ability be tested, it may be that test-taking strategies were dominated by a particular mode of thinking into which the student was 'set' early on in the test. This would certainly account for the finding of unidimensionality of the whole test. It is, however, a matter which requires further experimental research. If it turns out to be the case that the underlying 'dimension' is in fact some sort of test-taking mind-set, having little if anything to do with EFL reading or language, then our conclusions as to the EFL reading construct will be considerably weakened.

It also needs to be said that items of this kind should be compared with other EFL activities (such as listening) to see if the 'dimension' which appears to underly test performance is related to something as specific as reading or whether it is a more general language characteristic. Again, further research would be needed. There is, furthermore, a need to see if analyses of the kind presented here go beyond EFL to other languages, though this would be a complex methodological undertaking.

The sampling of the course content which was undertaken in order to provide test items may have been an inappropriate or inadequate procedure. The most that can be said for this is that it is at least an attempt to be systematic. It will, however, not be possible given the current state of our knowledge to have great confidence in sampling methods of any kind applied to data of this type.

While the use of the Rasch model in analysis appears to have been adequate, it may be that more realistic results would be obtained using a more sophisticated IRT model. While the arguments in favour of the Rasch model were presented in a favourable light in Chapter 5, it would still need to be demonstrated for data of this

type that other models were not superior. This issue was not addressed in the current study, largely through unavailability of the necessary computer programs. This is also an issue which tends to draw on arguments and procedures of a statistical kind rather than of a directly linguistic kind, and to that extent we continue to uphold our interdisciplinary tradition by relying on the judgments of those working in other 'universes of discourse'.

In connection with the working of the model, it may also be the case that the numbers involved here were too small for us to draw positive conclusions. It seems reasonable to suppose, however, that if consistency appears to be demonstrated in relatively small-scale analysis, then that same consistency will be demonstrated in larger applications. This is partly to raise the question highlighted earlier of extension to other groups, but the argument may also apply to the limited population of which the current group is a subset.

Another technical question which needs to be addressed is the appropriacy of factor analysis as a technique for investigating dimensionality. The problem arises when factor analytic methods are used to establish dimensionality (a necessary and unavoidable procedure) but are then extended to say something about the construct under investigation. We have at various places suggested that there are a number of shortcomings with factor analysis as a means of investigation the structure of proficiency, though it is a powerful tool which the researcher should be reluctant to renounce lightly. In the current study we have used factor analysis for the twin purposes of establishing dimensionality (a statistical concept) and of investigating construct; this may not be appropriate in the strictest sense, though the insights such an analysis offers seem to be of great benefit in attempting to make sense of test results.

Finally, the item types used in the pilot tests may be criticised for being of a too rigidly multiple choice format. Again, it may be that the tests were tests of ability to answer multiple choice questions rather than ability to read in EFL. This criticism is accepted, though we repeat that in the circumstances in which testing takes place, it is impossible to distinguish text and task and that all tests are to a greater or lesser extent 'impure' in that they interfere with the reader's interaction with text.

8.3. Dimensions of EFL reading

It is perhaps somewhat unsatisfactory to answer the question "What is reading?" by saying that it all depends on what you mean by reading. However, this is of particular relevance to EFL reading. Some people will take reading to mean the simple

mode in which language is realized or practised, while others will take it to represent a construct which exists in its own right, to be distinguished from, say, reading in a first language or from listening in a foreign language. Educational definitions and objectives will also be a significant part of any attempt to understand what we mean by reading in a foreign language. From one perspective, it is quite reasonable to claim that EFL reading is unidimensional since it describes a mode of activity rather than any psychological reality.

Using factor analysis, empirical support was found for the claim that EFL reading is unidimensional. No evidence was found for separate 'subskills' in comprehension, or indeed for the necessity to distinguish between comprehension and more limited 'reading' such as recognition of correct grammar.

The perspective and purpose of the teacher or the test designer will be an important feature in assessing the dimensionality of a set of test items. The absolute nature of proficiency in a foreign language seems to be open to serious doubt.

8.4. 'Ability' in EFL reading

By defining ability in terms of what testees can or cannot do, one is committed to a philosophy of criterion-referenced test construction, which may be difficult to apply to EFL reading in that such a method of test construction tends to be knowledge-bound. It also has the disadvantage of requiring exhaustive analysis of the domain, which may have the effect of creating fragments of the construct which it is tempting to see as additive or hierarchical in some way.

There lurks a conflict between criterion-referencing, which claims to be able to identify separate behaviours, and latent trait testing, which claims to account for test behaviour in terms of one or more essentially unknown traits (they wouldn't be 'latent' otherwise). This is a debate which will never leave us (it is the Ebel *versus* Horne debate which was adumbrated in Chapter 3), and which partly depends on the inclinations of the individual.

However, using criterion-referenced methods of test construction has the advantage of refining and crystallising thought about content at the test construction stage, which means that prior validity is at least more likely to be achieved this way. Moreover, in the present study, ability as a latent trait estimate was found to be stable in a fairly limited context; this now needs to be extended to other groups and also longitudinally - will those identified as 'able' prove to be so? This predictive validity could not be established within the time scale of the present study.

8.5. 'Difficulty' in EFL reading

Criterion-referencing as a method of test construction has the added advantage of highlighting what is and is not thought to be difficult about the construct under investigation. The problem is that this tends to encourage the idea that hierarchies will be found, when this may be conceptually misplaced in educational research (Horne 1984).

The present study found no evidence for specifically identifiable sources of difficulty in EFL reading, whether at the extremely limited level of individual grammar items, or at the more diffuse level of reading comprehension. Indeed, we found little evidence for a consistently rising scale of difficulty across separate prior 'dimensions'. This tends to confirm the emphasis we have placed on difficulty being a function of text and task together. There was slight evidence for a 'learning' factor, though this could also be some sort of 'test-taking' factor.

As far as the use of item-writing technologies is concerned, it has to be said that these suffer from being open to criticism on all the grounds on which criterion-referenced testing can be criticised. In particular they tend to encourage a fragmentation which may be inappropriate, and demand an exhaustiveness which is not only incompatible with the current state of our knowledge in most educational fields but is also extremely time-consuming. In favour of item-writing technologies it can be said that they do require systematic development of items, and if so viewed they have a positive contribution to make to test development. However, it may be unfortunate to view such methods as 'technologies' (implying objectivity and scientific accuracy), when in practice they are merely the codification of common sense.

8.6. Latent trait models

The general objections to latent trait models will not be repeated here, though it remains true that it is partly the inclination of the individual which determines whether he or she will look favourably upon the field. In particular, concern for 'humanistic' developments in education is likely to militate against any strong application of IRT models. The demands of test consumers may, however, ensure a continuing interest in the development of 'objective' models of measurement.

Empirically we have found that the Rasch model fits EFL reading data of the type used in the current analysis.

However, when the population is more or less as large as the sample upon which items are piloted, then there seems to be little advantage in using an IRT model.

After all, it was found that classical test statistics remain equally stable for this group.

Practical difficulties of interpretation remain for the test consumer. If test scores have to be relayed to the outside world then much thought needs to go into the kind of scale that will be used. Classical statistics have the virtue of being relatively easily interpretable.

There remains, also, the problem of labelling items in an item bank. This is not an issue which will be solved by test theory, but by, in this case, applied linguistic theory. Again, test purpose becomes of prime importance.

8.7. The future

Practical developments in item banking for EFL should now be pursued. In particular it will be important to see whether adaptive testing can be implemented in any useful way.

Further empirical work on the 'sample-freeness' of IRT statistics should give us significant insights into the structure of foreign language proficiency, though this will need to go hand in hand with more 'philosophical' consideration of what it is that we want to mean by the various constructs with which we work. This is an educational rather than a statistical issue.

BIBLIOGRAPHY

Abbott, M. M. (1972) *Publisher's management problems when entering into a new field of test development* in Curtis (ed.)

Adams, R.J., Griffin, P.E. and Martin,L. (1987) 'A latent trait method for measuring a dimension in second language proficiency.' *Language Testing* 3: 9-27

Ahmad, K., Corbett, G., Rogers, M. and Sussex, R. (1985) *Computers, Language Learning and Language Teaching* Cambridge: CUP

Alderson, J.C. (1981a) *Report of the discussion on Communicative Language Testing* in Alderson and Hughes (eds.)

Alderson, J.C. (1981b) *Report of the discussion on Testing English for Specific Purposes* in Alderson and Hughes (eds.)

Alderson, J.C. (1983) *Who needs jam?* in Hughes and Porter (eds.)

Alderson, J.C. (1984) *Reading in a Foreign Language: a reading problem or a language problem?* in Alderson and Urquhart (eds.)

Alderson, J.C. and Hughes, A. (eds.) (1981) *Issues in Language Testing* British Council: ELT documents 111

Alderson, J.C. and Urquhart, A.H. (1983) *The effect of student background discipline on comprehension: a pilot study* in Hughes and Porter (eds.)

Alderson, J.C. and Urquhart, A.H. (1984a) *What is Reading?* in Alderson and Urquhart (eds.)

Alderson, J.C. and Urquhart, A.H. (eds.) (1984b) *Reading in a Foreign Language* London: Longman

Alderson, J.C. and Urquhart, A.H. (1985) 'The effect of students' academic discipline on their performance on ESP reading tests' *Language Testing* 2.2: 192-204

Alexander, R. (1980) 'A learning-to-learn perspective on reading in a foreign language' *System* 8: 113-119

Allen, J.P.B. and Corder, S.P. (eds.) (1974) *The Edinburgh Course in Applied Linguistics. Vol.3: Techniques in Applied Linguistics* Oxford: OUP

Anastasi, A. (1983) *Traits, States, and Situations: A Comprehensive View* in Wainer and Messick (eds.)

Andersen, E.B. (1973) *Conditional Inference and Models for Measuring* Copenhagen: Mentalhygiejnisk Forlag

Anderson, J., Kearney, G.E. & Everett, A.V. (1968) 'An evaluation of Rasch's structural model for test items.' *British Journal of Mathematical and Statistical Psychology* 21: 231- 238

Anderson, R.C. (1972) 'How to construct achievement tests to assess comprehension' *Review of Educational Research* 42.2: 145-170

Anderson, R.C., Reynolds, R.E., Schallert, D.L. and Goetz, E.T. (1977) 'Frameworks for comprehending discourse' *American Educational Research Journal* 14.4: 367-381

Anderson, R.C. and Shifrin, Z. (1980) *The Meaning of Words in Context* in Spiro *et al.* (eds.)

Angoff, W.H. (1982) *Summary and derivation of equating methods used at ETS* in P.W. Holland & D.B. Rubin (eds.) *Test Equating* New York: Academic Press

Angoff, W.H. (1984) *Scales, norms and equivalent scores* Princeton NJ: Educational Testing Service

Ansfield, P.J. (1973) 'A User Oriented Computing Procedure for Compiling and Generating Examinations' *Educational Technology* 13.3: 12-13

Ard, J. (1985) 'Vantage points for the analysis of scientific discourse' *The ESP Journal* 4.3: 3-19

Ariew, R. (1979) 'A diagnostic test for students entering a computer- assisted learning curriculum in French' *Computers and Education* 3. 331-333

Ariew, R. (1982) 'A management system for foreign language tests' *Computers and Education* 6: 117-120

Atkinson, R.C. (1972) 'Ingredients for a theory of instruction' *American Psychologist* 27: 921-931

Baker, F.B.(1973) 'An interactive approach to test construction' *Educational Technology* 13.3: 13–15

Baker, F.B. (1974) *The Role of Statistics* in Lippey (ed.)

Baker, G.P. and Hacker, P.M.S. (1984) *Language, Sense and Nonsense* Oxford: Basil Blackwell

Baker, R. (1982) 'Measures of syntactic complexity in ESL composition' Unpublished M.Sc. thesis: University of Edinburgh

Baker, R. (1987) 'An investigation of the Rasch model in its application to foreign language proficiency testing' Unpublished Ph.D. dissertation: University of Edinburgh

Baltra, A. (1983) 'Learning how to cope with reading in English for academic purposes in 26 hours' *Reading in a Foreign Language* 1: 20–34

Barik, H. and Swain, M. (1975) *Three-year evaluation of a large-scale early grade French immersion program: The Ottawa Study* Monographs in Fundamental Education 8 Paris: Unesco 77–86

Barrett, T.C. (1968) *What is Reading* in T.Clymer (ed.) *Innovation and Change in Reading Instruction* (67th Year Book of the National Society for the Study of Education) University of Chicago Press

Barson, J., Smith, R., Levine, D., Scholl, M. and Scholl, P. (1981) *University-level CAI in French* in Suppes (ed.)

Bartlett, F.C. (1932) *Remembering* Cambridge: Cambridge University Press

Baten, L. and Cornu, A. -M. (1984) *Reading Stragegies for LSP texts: a theoretical outline on the basis of text function, with practical application* in Pugh and Ulijn (eds.)

Beaugrande, R. de (1984) *Reading Skills for Foreign Languages: a processing approach* in Pugh and Ulijn (eds.)

Bejar, I.I. (1980) 'A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates' *Journal of Educational Measurement* 17.4: 283–296

Bejar, I.I. (1983) *Achievement Testing: Recent Advances* Beverley Hills: Sage Publications

Bennett, W.A. (1974) *Applied Linguistics and Language Learning* London: Hutchinson

Bensoussan, M. (1982) 'Testing the test of advanced EFL reading comprehension: to what extent does the difficulty of a multiple-choice comprehension test reflect the difficulty of the text?' *System* 10.3: 285-90

Bensoussan, M., Goldenblatt, L. and Kreindler, I. (1984) 'Changing the difficulty level of multiple-choice EFL reading comprehension questions' *Language Testing* 1.1: 105-109

Berk, R.A. (1978) 'The application of structural facet theory to achievement test construction' *Educational Research Quarterly* 3: 62-72

Berk, R.A. (1980) *Item Analysis* in Berk (ed.)

Berk, R.A. (ed.) (1980) *Criterion-Referenced Measurement: The State of the Art* Baltimore: The John Hopkins University Press

Berkoff, N.A. (1979) 'Reading skills in extended discourse in English as a foreign language' *Journal of Research in Reading* 2.2: 95-107

Berman, R.A. (1984) *Syntactic components of the foreign language reading process* in Alderson and Urquhart (eds.)

Birnbaum, A. (1968) *Some latent trait models and their use in inferring an examinee's ability* in Lord and Novick (eds.)

Black, H.D. and Dockrell, W.B. (1980) *Diagnostic Assessment in Secondary Schools* The Scottish Council for Research in Education

Black, H.D. and Dockrell, W.B. (1984) *Criterion-Referenced Assessment in the classroom* Edinburgh; The Scottish Council for Research in Education

Blanton, L.L. (1984) 'Using a hierarchical model to teach academic reading to advanced ESL students: how to make a long story short' *The ESP Journal* 3: 37-46

Bloom, B.S. (1976) *Human Characteristics and School Learning* New York: McGraw Hill

Borich, G.D. and Jemelka, R.P. (1981) *Evaluation* in H.F.O'Neil (jnr) *Computer-based Instruction: A State of the Art Assessment* New York: Academic Press

Bormuth, J.R. (1969) *An Operational Definition of Comprehension Instruction* in

Goodman and Fleming (eds.)

Bormuth, J.R. (1970) *On The Theory of Achievement Test Items* Chicago: The University of Chicago Press

Bowker, D.C. (1984) 'The information gap in placement testing' *ELTJ* 38: 248–255

Boyd, G., Keller, A. and Kenner, R. (1982) 'Remedial and second language teaching using computer assisted learning' *Computers and Education* 6: 105–112

Boyle, T.A., Smith, W.F. and Eckert, R.G. (1976) 'Computer-mediated testing: a branched program achievement test' *The Modern Language Journal* 60: 428–440

Bransford, J.D., Stein, B.S. and Shelton, T. (1984) *Learning from the perspective of the comprehender* in Alderson and Urquhart (eds.)

Braun, T.E.D. and Mulford, G.W. (1984) 'Computer-assisted instruction as an integral part of a first-semester French curriculum' *Computers and the Humanities* 18: 47–56

Brebner, A., Johnson, K. and Mydlarski, D. (1984) 'CAI and second language learning: an evaluation of programs for drill and practice in written French' *Computers and Education* 8.4: 471–474

Breen, M.P. (1985) 'Authenticity in the language classroom' *Applied Linguistics* 6.1: 60–70

Brent, A. (1978) *Philosophical Foundations for the Curriculum* London: George Allen and Unwin

Brown, G. and Yule, G. (1983) *Discourse Analysis* Cambridge: CUP

Brown, H.I. & Holland, P.W. (1982) *Observed score test equating: a mathematical analysis of some ETS equating procedures* in Holland & Rubin (eds)

Brown, J.D. (1984) *A norm-referenced engineering reading test* in Pugh and Ulijn (eds.)

Brown, S. (1981) *What do they know? A review of criterion-referenced assessment* Edinburgh: HMSO

Brumfit, C.J. (1978) *The teaching of advanced reading skills in a foreign language with particular reference to English as a foreign language* in V. Kinsella (ed.) *Language Teaching and Linguistics Surveys* Cambridge: CUP

Brumfit, C., Phillips, M. and Skehan, P. (1985) *Computers in English Language Teaching* ELT Documents 122 Oxford: Pergamon Press for the British Council

Bruton, C. (1985) *The English Language Testing Service* Paper given to the Applied Linguistics department of Edinburgh University November 1985

Buckley-Sharp, M.D. (1973) 'A multiple-choice question banking system' *Educational Technology* 13.3: 16-18

Buckley-Sharp, M.D. and Harris, F.T.C. (1970a) 'A computer program for banking multiple choice questions' *The Computer Journal* 13.3: 230-236

Buckley-Sharp, M.D. and Harris, F.T.C. (1970b) 'The banking of multiple choice questions' *British Journal of Medical Education* 4: 42-52

Bung, K. (1973) *Towards a Theory of Programmed Learning Instruction* Mouton, The Hague: Janua Linguarum

Burnett, J.D. and Miller, L. (1984) 'Computer-assisted learning and reading: developing the product or fostering the process?' *Computers and Education* 8.1: 145-150

Byrne, C.J. (1976) 'Computerized question banking systems: I - The State of the Art' *British Journal of Educational Technology* 7.2: 44-64

Cambridge (1982) *Cambridge Examinations in English: Changes of Syllabus in 1984* Cambridge: University of Cambridge Local Examinations Syndicate

Campbell, D.T. and Fiske, D.W. (1969) 'Convergent and discriminant validation by the multi-trait, multi-method matrix' *Psychological Bulletin* 56.2: 81-105

Campen, J.V. (1981a) *A Computer-Assisted Course in Russian* in Suppes (ed.)

Campen, J.V. (1981b) *Computer-Generated Drills in Second Language Instruction* in Suppes (ed.)

Campen, J.V. (1981c) *A Computer-Assisted Introduction to the morphology of Old Church Slavic* in Suppes (ed.)

Campen, J.V., Markosian, L.Z. and Seropian, H. (1981) *A Computer-Assisted Language Instruction System with initial application to Armenian* in Suppes (ed.)

Canale, M. and Swain, M. (1980) 'Theoretical bases of communicative approaches to

second language teaching and testing' *Applied Linguistics* 1: 1–47

Carnine, D. and Silbert, J. (1979) *Direct Instruction Reading* Columbus, Ohio: Merrill

Carpenter, P.A. (1984) *The influence of methodologies on psycholinguistic research: a regression to the Whorfian hypothesis* in Kieras and Just (eds.)

Carrier, M. (1985) 'Computer-assisted language learning review' *ELTJ* 39.2: 131–134

Carroll, B.J. (1980) *Testing Communicative Performance* Oxford: Pergamon Press

Carroll, B.J. (1981) *Specifications for the English Language Testing Service* in Alderson and Hughes (eds.)

Carroll, J.B. (1970) *The Nature of the Reading Process* in Gunderson (ed.)

Carroll, J.B. (1972) *Defining Language Comprehension: Some Speculations* in Freedle and Carroll (eds.)

Carver, R.P. (1973) *Reading as reasoning: implications for measurement* in W.H.MacGintie (ed.) *Assessment Problems in Reading* Newark, Delaware: International Reading Association

Carver, R.P. (1978) 'The case against statistical significance testing' *Harvard Educational Review* 48.3: 378–399

Cattell, R.B. (1965) 'Factor analysis: an introduction to essentials. (I) the purpose and underlying models, (II) the role of factor analysis in research.' *Biometrics* 21: 190–215, 405–435

Catterson, J. (1979) *Comprehension: the argument for a discourse analysis model* in Pennock (ed.)

Chandler, D. (ed.) (1983) *Exploring English with Microcomputers* London: Council for Education Technology

Chappelle, C. and Jamieson, J. (1984) *Language Lessons on the Plato IV System* in Wyatt (ed.)

Charniak, E. (1981) 'A common representation for problem-solving and language comprehension' *Artificial Intelligence* 16: 225–255

Chen, Z. and Henning, G. (1985) 'Linguistic and cultural bias in language proficiency tests' *Language Testing* 2.2: 155–163

Childs, R. (1978) *Item Banking* Slough: National Foundation for Educational Research [Basic Testing Series]

Chomsky, N. (1976) *Reflections on Language* Glasgow: Collins/Fontana

Choppin, B.H. (1974) *Rasch/Choppin Pair-Wise Analysis: Express Calibration by Pair-X* Slough: National Foundation for Educational Research

Choppin, B.H. (1976) *Recent developments in item banking: a review* in de Gruijter and van der Kamp (eds.)

Choppin, B.H. (1978) *Item Banking and the Monitoring of Achievement* Slough: National Foundation for Educational Research

Choppin, B.H. (1979) 'Testing the questions: the Rasch model and item banking' in Raggett *et al.* (eds.)

Choppin, B.H. (1981) *Educational measurement and the item bank model* in Lacey and Lawton (eds.)

Choppin, B.H. and Orr, L. (1976) *Aptitude Testing at Eighteen Plus* Slough: National Foundation for Educational Research

Cito (1985) 'Plans for research' *language Testing Notes* December 1985

Clarke, M.A. (1979) 'Reading in Spanish and English: evidence from adult ESL students' *Language Learning* 29.1: 121–147

Clarke, M.A. and Silberstein, S. (1979) 'Toward a realization of psycholinguistic principles in the ESL reading class' in Mackay *et al.* (eds.)

Clymer, T. (1968) *What is Reading? Some current concepts* in A.M. Robinson (ed.) *Innovation and Change in Reading Instruction* Chicago: University of Chicago Press

Clymer, T. (1969) *Behavioral Objectives for Reading* Boston,Mass.: Grinn and Co.

Coady, J. (1979) *A Psycholinguistic Model of the ESL Reader* in Mackay *et al.* (eds.)

Cohen, A. (1980) *Testing Language Ability in the Classroom* Rowley, Mass.: Newbury House

Cohen, A.D., Glasman, H., Rosenbaum, P.R., Ferrara, J. & Fine, J. (1979) 'Reading English for specialized purposes: discourse analysis and the use of students' *TESOL Quarterly* 13: 551 - 564

Cohen, G. (1979) 'Language comprehension in old age' *Cognitive Psychology* 11: 412-429

Cole, P., Lebowitz, R. and Hart, R. (1984) 'Teaching Hebrew with the the aid of computers: The Illinois Program' *Computers and the Humanities* 18: 87-99

Coles, G.S. (1978) 'The learning disabilities test battery: empirical and social issues' *Harvard Educational Review* 48: 313-340

Collett, M.J. (1980) 'Examples of applications of computers to modern language study. 1 The step wise development of programs in reading, grammar and vocabulary' *System* 8: 195-204

Conoley, J.C. and O'Neil, H.F. (1979) *A Primer For Developing Test Items* in O'Neil (ed.)

Cook, L.L., Eignor, D.R. & Taft, H.L. (1984) *A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates* Paper presented at the annual meeting of the American Educational Research Association, New Orleans

Cooper, M. (1984) *Linguistic competence of practised and unpractised non-native readers of English* in Alderson and Urquhart (eds.)

Corrick, M. (1984) 'Something to be clearly understood' *Guardian* 19th June 1984

Cotton, J.W., Gallagher, J.P. and Marshall, S.P. (1977) 'The identification and decomposition of hierarchical tasks' *American Educational Research Journal* 14.3: 189-212

Cowell, W. (1981) 'Applicability of a simplified three-parameter logistic model for equating tests' Paper presented at the annual meeting of the American Educational Research Association, New Orleans

Criper, C. (1981) *Reaction to the Carroll paper - 2* in Alderson and Hughes (eds.)

Criper, C. and Davies, A. (1986) *Edinburgh ELTS Validation Project: Project Report* Edinburgh University

Cronbach, L.J. (1951) 'Coefficient alpha and the internal structure of tests' *Psychometrika* 16: 297–334

Cronbach, L.J. (1970) Review of Bormuth (1970) *Psychometrika* 35.4: 509–11

Cronbach, L.J. and Gleser, G.C. (1965) *Psychological Tests and Personnel Decisions* Urbana: University of Illinois Press

Cronbach, L.J., Gleser, G.C., Nanda, H. and Rajaratnam, N. (1972) *The Dependability of Behavioral Measurements* New York: John Wiley and Sons

Cronbach, L.J. and Meehl, P.E. (1955) 'Content validity in psychological tests' *Psychological Bulletin* 52.4: 281–302)

Crookes, G. (1986) 'Towards a validated analysis of scientific text structure' *Applied Linguistics* 7.1: 57–70

Culley, G.R. (1984) 'Generic or specific: having it both ways with generative CAI' *Computers and the Humanities* 18: 183–188

Cummins, J. (1979) *Cognitive/academic language proficiency, linguistic interdependence, the optimal age question and some other matters* Working Papers on Bilingualism 19: 197–205)

Cummins, J. (1980) 'The cross-lingual dimensions of language proficiency: implications for bilingual education and the optimal age issue' *TESOL Quarterly* 14: 175–187

Curtis, H.A. (1972) *Commercially produced item banks: the local project director's responsibilities* in Curtis (ed.)

Curtis, H.A. (ed.) (1972) *The Development and Management of Banks of Performance Based Test Items* New York: Harcourt, Brace, Jovanovich [ERIC document no. ED 072099]

Cziko, G. (1970) 'Reading in a second language: linguistic constraints' *Language Learning* 30.1: 101–116

Cziko, G.A. (1978) 'Differences in first and second language reading: the use of syntactic, semantic and discourse constraints' *The Canadian Modern Language Review* 34.3: 473–489

Cziko, G.A. and Lin, N.-H.J. (1984) 'The construction and analysis of short scales of language proficiency: classical psychometric, latent trait, and nonparametric approaches' *TESOL Quarterly* 18.4: 627-647

Dakin, J. (1973) *The Language Laboratory and Language Learning* London: Longman

Dale, E. and Chall, J. (1948) 'A Formula for predicting readability' *Educational Research Bulletin* 27: 11-20, 28

Davies, A. (1977) *The Construction of Language Tests* in J.P.B. Allen and A.Davies (eds.) *The Edinburgh Course in Applied Linguistics Volume 4: Testing and Experimental Methods* Oxford: OUP

Davies, A. (1982) *Criteria for evaluation of tests of EFL* in J.B. Heaton (ed.) *Language Testing* Modern English Publications

Davies, A. (1982) *Language Testing* in Kinsella (ed.)

Davies, A. (1983) *The validity of concurrent validation* in Hughes and Porter (eds.)

Davies, A. and Widdowson, H.G. (1974) *Reading and Writing* in Allen and Corder (eds.)

Davis, F.B. (1944) 'Fundamental factors of comprehension in reading' *Psychometrika* 9: 185-197

Davis, F.B. (1946) 'Comment on Thurstone' *Psychometrika* 11: 249-255

Davis, F.B. (1968) 'Research in comprehension in reading' *Reading Research Quarterly* 4: 499-545

Davis, F.B. (1972) 'Psychometric research on comprehension in reading *Reading Research Quarterly* 7: 628-678

Davison, A. and Kantor, R.N. (1982) 'On the failure of readability formulas to define readable texts: a case study from adaptations' *Reading Research Quarterly* 17: 187-209

de Gruijter, D.N.M. and van der Kamp, L.J.Th. (eds.) (1976) *Advances in Psychological and Educational Measurement* London: John Wiley and Sons

Demaiziere, F. (1982) 'An experiment in computer-assisted learning of English grammar at the University of Paris VII' *Computers and Education* 6: 121-125

Denney, C. (1973) 'There is more to a pool than data collection' *Educational Technology* 13.3: 19–20

Deyes, T. (1984) 'Towards an authentic "discourse cloze"' *Applied Linguistics* 5.2: 128–137

Diederich, P.B. (1970) Review of Bormuth (1970) *Educational and Psychological Measurement* 30: 1003–5

Dieterich, T., Freeman, C. and Griffin, P. (1978) *Assessing Comprehension in a School Setting* [Papers in Applied Linguistics; Linguistics and Reading Series: 3] Arlington,Virginia: Center for Applied Linguistics

Divgi, D.R. (1981) 'Does the Rasch model really work? Not if you look closely' Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Doerr, C. (1979) *Microcomputers and the 3 R's: a guide for teachers* Rochelle Park, N.J.: Hayden Book Co.

Douglas, D. (1978) 'Gain in reading proficiency in English as a foreign language measured by three cloze scoring methods' *Journal of Research in Reading* 1.1: 67–73)

Drasgow, F. and Parsons, C.K. (1983) 'Application of unidimensional item response models to multidimensional data' *Applied Psychological Measurement* in press

Drever, E. (1983) 'Curriculum objectives as assessment criteria – some problems of validity' *Programmed Learning and Educational Technology* 20.1: 54–57

Duchastell, P.C. and Merrill, P.F. (1973) 'The effects of behavioral objectives on learning: a review of empirical studies' *Review of Educational research* 43: 53–70

Duell, O.K. (1974) 'Effect of type of objective, level of test questions, and the judged importance of tested materials upon post-test performance' *Journal of Educational Psychology* 66: 225 – 232

Ebel, R.L. (1961) 'Must all tests be valid?' *American Psychologist* 16.10: 640–647

Ebel, R.L. (1966) 'The value of internal consistency in classroom examinations' *Journal of Educational Measurement* 5: 71–74

Eisner, E.W. (1979) *The Educational Imagination* New York: Macmillan

Embretson, S. (1984) 'A general latent trait model for response processes' *Psychometrika* 49.2: 175–186

Eskey, D.E. (1973) 'A model program for teaching advanced reading to students of English as a foreign language' *Language Learning* 23.2: 169–184

Eskey, D.E. (1979) *A model program for teaching advanced reading to students of EFL* in Mackay *et al.* (eds.)

Farr, R. (1969) *Reading: What Can Be Measured?* Newark, Delaware: International Reading Association

Farrington, B. (1982) 'Computer-based exercises for language-learning at university level' *Computers and Education* 6: 113–116

Feitelson, D. (ed.) *Mother Tongue or Second Language? On the teaching of reading in multilingual societies* Newark, Delaware: International Reading Association

Ferguson, G.A. (1949) 'On the theory of test discrimination' *Psychometrika* 14.1: 61–68

Ferraris, M., Midoro, V. and Olimpo, G. (1984) 'Diagnostic testing and the development of CAL remedial sequences' *Computers and Education* 8.4: 407–414

Fillmore, C.J. (1982) *Ideal Readers and Real Readers* in Tannen (ed.)

Filstead, W.J. (ed.) (1970) *Qualitative Methodology: firsthand involvement with the social world* Chicago: Markham Publishing Co.

Finn, P.J. (1975) 'A question-writing algorithm' *Journal of Reading Behaviour* 7: 341–367

Fischer, G.H. (1974) *Einfuhrung in die Theorie psychologischer Tests. Grundlagen und Anwendungen* Bern: Huber

Fischer, G.H. (1978) 'Probabilistic test models and their applications' *German Journal of Psychology* 2: 298–319

Fischer, G.H. and Kisser, R. (1983) *Notes on the Exponential Latency Model and an Empirical Application* in Wainer and Messick (eds.)

Fischer, G.H. and Pendl, P. (1976) *Individualised testing on the basis of the dichotomous Rasch model* in de Gruijter and van der Kamp (eds.)

Flahive, D.E. (1980) *Separating the g factor from reading comprehension* in Oller and Perkins (eds.)

Flanagan, J.C. (1982) *Discussion of "Some issues in test equating" in Holland & Rubin (eds*

Forsyth, R., Saisangjan, U. & Gilmer, J. (1981) 'Some empirical results related to the robustness of the Rasch model' *Applied Psychological Measurement* 5: 175–186

Fransson, A. (1984) *Cramming or understanding? Effects of intrinsic and extrinsic motivation on approach to learning and test performance*

Frederiksen, C.H. (1972) *Effects of task-induced cognitive operations on comprehension and memory processes* in Freedle and Carroll (eds.)

Freebody, P. and Anderson, R.C. (1983) 'Effects of vocabulary difficulty text cohesion and schema availability on reading comprehension' *Reading Research Quarterly* 18.3: 277–294

Freedle, R.O. and Carroll, J.B. (1972) *Language Comprehension and the Acquisition of Knowledge* New York: John Wiley and Sons

Freedle, R.O. and Carroll, J.B. (1972) *Language Comprehension and the Acquisition of Knowledge: Reflections* in Freedle and Carroll (eds.)

Fremer, J.J. and Anastasio, E.J. (1969) 'Computer-assisted item-writing: I. Spelling Items' *Journal of Educational Measurement* 6.2: 69–74

French, S. (1981) 'Measurement theory and examinations' *British Journal of Mathematical and Statistical Psychology* 34: 38–49

Gagne, R.M. (1962) 'The acquisition of knowledge' *Psychological Review* 69.4: 355–365

Gagne, R.M. (1970) *The Conditions of Learning* New York: Rinehart and Winston

Gibson, E.J. & Levin, H. (1975) *The psychology of reading* Cambridge, Mass.: MIT Press

Goetz, E.T. and Armbruster, B.B. (1980) *Psychological Correlates of Text Structure* in Spiro *et al.* (eds.)

Goldstein, H. (1979) 'Consequences of using the Rasch model for educational

assessment' *British Educational Research Journal* 5.2: 211–220

Goldstein, H. (1980) 'Dimensionality, bias, independence and measurement scale problems in latent trait test score models' *British Journal of Mathematical and Statistical Psychology* 33: 234–246

Goldstein, H. (1981) *Limitations of the Rasch model for educational assessment* in Lacey and Lawton (eds.)

Goldstein, H. and Blinkhorn, S. (1977) 'Doubts about item banking' *Bulletin of the British Psychological Society* 30: 309–311

Goodacre, E. (1979) 'What is reading: which model?' in Raggett *et al.* (eds.)

Goodman, K.S. (1967) 'Reading: a psycholinguistic guessing game' *reprinted* in Goodman (1982)

Goodman, K.S. (1968) *The Psycholinguistic Nature of the Reading Process* Detroit: Wayne University Press

Goodman, K.S. (1970) 'Psycholinguistic Universals in the reading process' *Journal of Typographical Research* 4: 103–110

Goodman, K.S. (1982) *Language and Literacy* (2 vols.) London: Routledge and Kegan Paul

Goodman, K.S. and Fleming, J.T. (eds.) *Psycholinguistics and the Teaching of Reading* Newark,Delaware: International Reading Association

Gorth, W.P., Allen, D.W. and Grayson, A. (1971) 'Computer programs for test objective and item banking' *Educational and Psychological Measurement* 31: 245–250

Gough, P.B. (1972) *One Second of Reading* in Kavanagh and Mattingley (eds.)

Graesser, A.C. and Black, J.B. (eds.) (1985) *The Psychology of Questions* Hillsdale, New Jersey: Lawrence Erlbaum Associates

Grellet, F. (1981) *Developing Reading Skills* Cambridge: CUP

Green, B.F. (1983) *The Promise of Tailored Tests* in Wainer and Messick (eds.)

Green, S.B., Lissitz, R.W. and Mulaik, S.A.(1977) 'Limitations of coefficient alpha as an index of test unidimensionality' *Educational and Psychological Measurement* 37:

827-838

Guilford, J.P. and Fruchter, B. (1981) *Fundamental Statistics in Psychology and Education* (6th Edition) Tokyo: McGraw-Hill

Gunderson, D.V. (ed.) (1970) *Language and Reading* Washington D.C.: Center for Applied Linguistics

Guskey, T.R. (1981) 'Comparison of a Rasch model scale and the grade- equivalent scale for vertical equating of test scores.' *Applied Psychological Measurement* 5: 187-201

Gustafsson, J.-E. (1977) *The Rasch model for dichotomous items: theory applications and a computer program* Reports from the Institute of Education University of Goteburg No. 63 [ED 154018]

Gustafsson, J.-E. (1979) 'The Rasch model in vertical equating of tests: a critique of Slinde and Linn.' *Journal of Educational Measurement* 16: 153-158

Gustafsson, J.-E. (1980a) 'Testing and obtaining fit of data to the Rasch model' *British Journal of Mathematical and Statistical Psychology* 33: 205-233

Gustafsson, J.-E. (1980b) 'A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items' *Educational and Psychological Measurement* 40: 377-385

Gustafsson, J.-E. and Lindblad, T. (1978) *The Rasch model for dichotomous items: a solution of the conditional estimation problem for long tests and some thoughts about item screening procedures* Reports from the Institute of Education, University of Goteburg, No. 67

Guttman, L.L. (1950) *The Basis for Scalogram Analysis* in S.A. Stouffer *et al.* (eds.) *Measurement and Prediction* Princeton, New Jersey: Princeton University Press

Guttman, L. (1971) 'Measurement as structural theory' *Psychometrika* 36.4: 329-347

Guzzetti, B.J. (1984) 'The reading process in content fields: a psycholinguistic investigation' *American Educational Research Journal* 21.3: 659-668

Haertel, E.H. (1980) *Determining what is measured by multiple-choice tests of reading comprehension* Unpublished Ph.D. thesis, University of Chicago

Haertel, E.H. (1984) 'Detection of a skill dichotomy using standardized achievement test items' *Journal of Educational Measurement* 21.1: 59–72

Haertel, E. (1985) 'Construct validity and criterion-referenced testing.' *Review of Educational Research* 55.1: 23–46

Haksar, L. (1983) 'Design and usage of an item bank' *Programmed Learning and Educational Technology* 20.4: 253–262

Hall, D. (1983) *Review* of D.D. Sim and B. Laufer-Dvorkin *Reading Comprehension Course: Selected Strategies* London: Collins *Reading in a Foreign Language* 1.2: 138–140

Halle, M. and Stevens, K.N. (1964) 'Speech recognition: a model and a program for research' in J.A. Fodor and J.J. Katz (eds.) *The structure of language: readings in the philosophy of language* Englewood Cliffs, N.J.: Prentice Hall Inc.

Hambleton, R.K. (1980) *Test score validity and standard-setting methods* in Berk (ed.)

Hambleton, R.K. and Cook, L.L. (1977) 'Latent trait models and their use in the analysis of educational test data' *Journal of Educational and Psychological Measurement* 14.2: 75–96

Hambleton, R.K. and Swaminathan, H. (1985) *Item Response Theory* Boston: Kluwer-Nijhoff Publishing

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R. and Gifford, J.A. (1978) 'Developments in latent trait theory: models, technical issues and applications' *Review of Educational Research* 48: 467–510

Hambleton, R.K., Swaminathan, H., Algina, J. and Coulson, D.B. (1978) 'Criterion-referenced testing and measurement: a review of technical issues and developments' *Review of Educational Research* 48.1: 1–47

Harris, D.P. (1969) *Testing English as a Second Language* McGraw Hill

Harris, D.J. & Kolen, M.J. (1985) *Effect of examinee group on equating relationships* Paper presented at the annual meeting of the American Educational Research Association, Chicago

Harrison, A. (1983) *A Language Testing Handbook* London: Macmillan

Harrison, C. (1979) *Assessing the Readability of School Texts* in Lunzer and Gardner (eds.)

Harrison, C. and Dolan, T. (1979) *Reading comprehension - a psychological viewpoint* in Mackay *et al.* (eds.)

Hatch, E. (1974) 'Research on reading a second language' *Journal of Reading Behaviour* 6.1: 53-61

Hatch, E. and Farhady, H. (1982) *Research Design and Statistics for Applied Linguistics* Rowley,Mass.: Newbury House

Hatch, E., Polin, P. and Part, S. (1974) 'Acoustic scanning and syntactic processing: three reading experiments - first and second language learners' *Journal of Reading Behaviour* 6.3: 275-285

Hazlett, C.B. (1973) 'MEDSIRCH: multiple choice test items' *Educational Technology* 13.3: 24-26

Healy, J.M. (1982) 'The enigma of hyperlexia' *Reading Research Quarterly* 17.3: 319-338

Heaton, D.P. (1975) *Writing English Language Tests* London: Longman

Henning, G., Hudson, T. and Turner, J. (1985) 'Item response theory and the assumption of unidimensionality' *Language Testing* 2.2: 141-154

Herman, L.M., Richards, D.G. and Wolz, J.P. (1984) 'Comprehension of sentences by bottlenosed dolphins' *Cognition* 16.2: 129-219

Hewitt, G. (1982) 'A critique of research methods in the study of reading' *British Educational Research Journal* 8.1: 9-21

Higgins, J. and Johns, T. (1984) *Computers in Language Learning* London: Collins ELT

Hill, R.A. (1984) 'An investigation into morpheme acquisition studies with particular reference to a group of students in an 'acquisition-poor' environment.' Unpublished M.Sc. paper: University of Edinburgh

Hillocks, G.(Jr) and Ludlow, L.H. (1984) 'A taxonomy of skills in reading and interpreting fiction' *American Educational Research Journal* 21.1: 7-24

Hisama, K.K. (1980) *An analysis of various ESL proficiency tests* in Oller and Perkins (eds.)

Hively, W. (1974) 'Introduction to domain-referenced testing' *Educational Technology* 14: 5-10

Hively, W., Patterson, H.L. and Page, S. (1968) 'A universe-defined system of arithmetic achievement tests' *Journal of Educational Measurement* 5: 275-290

Hlynka, D. and Nelson, B. (1985) 'Educational technology as metaphor' *Programmed Learning and Educational Technology* 22.1: 7-15

Holland, P.W. and Rubin, D.B. (1983) *On Lord's Paradox* in Wainer and Messick (eds.)

Holmes, S.E. (1982) 'Unidimensionality and vertical equating with the Rasch model.' *Journal of Educational Measurement* 19: 139-147

Horn, J.L. (1966) 'Is it reasonable for assessments to have different psychometric properties than predictors?' *Journal of Educational Measurement* 5: 75-78

Horne, S.E. (1983) 'Learning hierarchies: a critique' *Educational Psychology* 3.1: 63-77

Horne, S.E. (1984) 'Criterion-referenced testing: pedagogical implications' *British Educational Research Journal* 10.2: 155-173

Hornke, L.F. and Sauter, M.P. (1981) *Testing English as a foreign language by an adaptive test strategy* in Allen James and Paul Westney (eds.) *New Linguistic Impulses in Foreign Language Teaching* Tubingen: Gunter Narr Verlag

Horton, T.R. (1973) *The Reading Standards of Children in Wales* Slough: National Foundation for Educational Research

Hsu, T.C. and Carlson, M. (1973) 'Test construction aspects of the computer assisted testing model' *Educational Technology* 13.3: 26-27

Hudson, T. (1982) 'The effects of induced schemata on the "short circuit" in L2 reading: non-decoding factors in L2 reading performance' *Language Learning* 32: 1-31

Hudson, T. and Lynch, B. (1984) 'A criterion-referenced measurement approach to ESL achievement testing' *Language Testing* 1.2: 171-201

Huey, E.B. (1908) *The Psychology and Pedagogy of Reading* Reprinted by MIT press

1968

Hughes, A. (1983) *Where now?* in Hughes and Porter (eds.)

Hughes, A. and Porter, D. (eds.) (1983) *Current Developments in Language Testing* London: Academic Press

Hulin, C.L., Drasgow, F. and Parsons, C.K. (1983) *Item Response Theory - Application to Psychological Measurement* Homewood, Illinois: Dow Jones - Irwin

Hullen, W. (1982) *Pedagogical considerations in teaching foreign languages at school* in Jung (ed.)

Hunt, K. (1966) 'Recent measures in syntactic development' in Lester, M. (ed.) *Readings in Applied Transformational Grammar*

Hunt, K. (1971) 'Teaching syntactic maturity' in Perren, G.E. and Trim, J.L.M. (eds.) *Applications of Linguistics* Cambridge: CUP

Hunter, J.E. and Schmidt, F.L. (1976) 'Critical analysis of the statistical and ethical implications of various definitions of test bias' *Psychological Bulletin* 83: 1053-1071

Hurley, P. and Hlynka, D. (1984) 'Prisoners of the cave: can instructional technology improve education?' *Computers and Education* 8.4: 427-434

Jackson, P. (1984) 'Towards a theory of topics' *Computers and Education* 8.1: 21-26

Jackson, P.M. & McLelland, R. (1979) *Artificial Intelligence* New York: Academic Press

Jenkinson, M.D. (1970) *Sources of knowledge for theories of reading* in Gunderson (ed.)

Johnson, S. and Maher, B. (1984) 'A thesaurus-linked science question-banking system' *British Journal of Educational Technology* 15.1: 14-23

Johnston, P. (1983) *Reading Comprehension Assessment* Newark,Delaware: International Reading Association

Johnston, P.H. (1984) *Assessment in Reading* in Pearson (ed.)

Jones, C. (1983) 'Computer-assisted language learning: testing or teaching?' *ELTJ* 37: 247-250

Jung, U.O.H. (ed.) (1982) *Reading: a symposium* Oxford: Pergamon

Kaplan, R. (1976) 'Effects of grouping and response characteristics of instructional objectives on learning from prose' *Journal of Educational Psychology* 68: 424-430

Kaplan, R. and Rothkopf, E.Z. (1974) 'Instructional objectives as directions to learners: effect of passage length and amount of objective-relevant content' *Journal of Educational Psychology* 66: 448-456

Kartunnen, L. (1970) *On The Semantics of Complement Sentences* in M.A. Campbell et al. (eds.) *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society* Chicago: Chicago Linguistic Society

Katz, J.J. (1973) 'On defining "presupposition"' *Linguistic Inquiry* 4: 256-60

Kavanagh, J.F. and Mattingley, I.G. (eds.) (1972) *Language by Ear and by Eye: The Relationships between Speech and Writing* Cambridge, Mass.: MIT Press

Keeney, R.L. and Raiffa, H. (1977) *Decisions with Multiple Objectives* New York: Wiley

Kelderman, H. (1984) 'Loglinear Rasch model tests' *Psychometrika* 49: 223-245

Kellerman, M. (1981) *The Forgotten Third Skill* Oxford: Pergamon

Kennedy, A. (1984) *The Psychology of Reading* London: Methuen

Kempa, R.F. and L'Odiaga, J. (1984) 'Criterion-referenced interpretation of examination grades' *Educational Research* s6.1: 56-64

Kenning, M.J. and Kenning, M.-M. (1983) *Introduction to Computer-Assisted Language Teaching* Oxford: OUP

Kidd, M.E. and Holmes, G. (1984) 'CAL evaluation: a cautionary word' *Computers and Education* 8.1: 77-84

Kieras, D.E. and Just, M.A. (eds.) *New Methods in Reading Comprehension Research* Hillsdale, New Jersey: Lawrence Erlbaum Associates

Kim, J.-O. and Mueller, C.W. (1978) *Factor analysis: statistical methods and practical issues* Beverly Hills: Sage Publications

Kinsella, V. (ed.) (1982) *Surveys I* Cambridge: CUP

Kintsch, W. & van Dijk, T.A. (1978) 'Towards a model of text comprehension and production' *Psychological Review* 85: 363–394

Klare, G.R. (1978) *Assessing Readability* in Chapman, L.J. and Czerniewska, P. (eds.) *Reading: from process to practice* London: Routledge and Kegan Paul

Klein-Braley, C. and Stevenson, D.K. (eds.) (1981) *Practice and Problems in Language Testing 1* Frankfurt: Verlag Peter D. Lang: Orbis Linguisticus

Kolen, M.J. (1981) 'Comparison of traditional and item response theory methods for equating tests.' *Journal of Educational Measurement* 18: 1–11

Kolen, M.J. & Whitney, D.R. (1982) 'Comparison of four procedures for equating the tests of General Educational Development.' *Journal of Educational Measurement* 19: 279–293

Kolers, P.A. (1969) *Reading is only incidentally visual* in Goodman and Fleming (eds.)

Krantz, D.H., Luce, R.D., Suppes, P. and Tuersky, A. (1971) *Foundations of Measurement. Vol.1* New York: Academic Press

Krashen, S.D. (1985) *The Input Hypothesis: Issues and Implications* New York: Longman

Kronik, J.W. (n.d.) 'Feijoo and the fabrication of Fortunata' in Goldman, P.B. (ed.) *Conflicting Realities: Four Readings of a Chapter by Perez Galdos* London: Tamesis Books

La Berge, D. and Samuels, S.J. (1976) *Towards a Theory of Automatic Information Processing in Reading* in Singer and Ruddell (eds.)

Lacey, C. and Lawton, D. (eds.) (1981) *Issues in Evaluation and Accountability* London: Methuen

Lado, R. (1961) *Language Testing* McGraw Hill

Laroche, J.M. (1979) 'Readability measurement for foreign-language materials' *System* 7: 131–135

Larsen-Freeman, D. and Strom, V. (1977) 'The construction of a second language acquisition index of development' *Language Learning* 27: 123–134

Leiblum, M.D. (1979) 'Screening for CAL' *Computers and Education* 3: 313–323

Leishman, J.B. (1956) *Translating Horace* Oxford: Bruno Cassirer

Levenston, E.A., Nir, R. and Blum-Kulka, S. (1984) *Discourse analysis and the testing of reading comprehension by cloze techniques* in Pugh and Ulijn (eds.)

Levin, H. (1970) *Reading Research: What, Why and For Whom?* in Gunderson (ed.)

Levine, A., Markosian, L.Z., Seropian, H. and Ferguson, C. (1981) *VERPS: a verb exercise and reference program with speech for Arabic language instruction* in Suppes (ed.)

Levine, D.R. (1981) *Computer-based analytic grading for German grammar instruction* in Suppes (ed.)

Libaw, F.B. (1973) 'Constructing tests with the MENTREX tutorial testing system' *Educational Technology* 13.3: 30–31

Libaw, F.B. (1974) *Pedagogical Implications* in Lippey (ed.)

Linn, R.L., Rock, D.A. and Cleary, T.A. (1969) 'The development and evaluation of several programmed testing methods' *Educational and Psychological Measurement* 29: 129–146

Lippey, G. (1973) 'The computer can support test construction in a variety of ways' *Educational Technology* 13.3: 10–12

Lippey, G. (ed.) (1974) *Computer-Assisted Test Construction* Englewood Cliffs, New Jersey: Educational Technology Publications

Lippey, G. (1974) *Overview* in Lippey (ed.)

Loevinger, J. (1947) *A systematic approach to the construction and evaluation of tests of ability* Psychological Monographs No. 61

Loevinger, J. (1965) 'Person and population as psychometric concepts' *Psychological Review* 72.2: 143–155

Lord, F.M. (1967) 'A paradox in the interpretation of group comparisons' *Psychological Bulletin* 68: 304–305

Lord, F.M. (1971) 'A theoretical study of the measurement effectiveness of flexilevel

tests' *Educational and Psychological Measurement* 31: 805–813

Lord, F.M. (1980) *Applications of Item Response Theory to Practical Testing Problems* Hillsdale, New Jersey: Lawrence Erlbaum Associates

Lord, F.M. and Novick, M.R. (1968) *Statistical Theories of Mental Test Scores* Reading, Massachussets: Addison-Wesley

Loret, P.G., Seder, A., Bianchini, J.C. & Vale, C. (1974) *Anchor Test study final report : Project report* vols 1–10. Berkeley, CA: Educational Testing Service

Lovett, M.W. (1981) *Reading skill and its development; theoretical and empirical considerations* in Mackinnon and Waller (eds.)

Loyd, B.H. and Hoover, H.D. (1980) 'Vertical equating using the Rasch model' *Journal of Educational Measurement* 17.3: 179–193

Lucas, P.A. and McConkie, G.W. (1980) 'The definition of test items: a descriptive approach' *American Educational Research Journal* 17.3: 133–140

Lumsden, J. (1961) 'The construction of unidimensional tests' *Psychological Bulletin* 58.2: 122–131

Lumsden, J. (1976) *Test Theory* in M.R. Rosenzweig and L.W. Porter (eds.) *Annual Review of Psychology, 27* Palo Alto: Annual Reviews Inc.

Lumsden, J. (1978) 'Tests are perfectly reliable' *British Journal of Mathematical and Statistical Psychology* 31: 19–26

Lunzer, E. and Gardner, K. (eds.) (1979) *The Effective Use of Reading* London: Heinemann (for the Schools Council)

Lunzer, E., Waite, M. and Dolan, T. (1979) *Comprehension and Comprehension Tests* in Lunzer and Gardner (eds.)

Mackay, R., Barkman, B. and Jordan, R.R. (eds.) (1979) *Reading in a Second Language: Hypotheses, Organisation and Practice* Rowley,Mass.: Newbury House

Mackay, R. and Mountford, A.J. (1978) *English for Specific Purposes* London: Longman

Mackay, R. and Mountford,A. (1979) *Reading for Information* in Mackay *et al.* (eds.)

Mackinnon, G.E. and Waller, T.G. (eds.) (1981) *Reading Research: Advances in Theory and Practice* Volume 3 New York: Academic Press

Mackworth, N.H. (1977) *The Line of Sight Approach* in Wanat (ed.) (1977b)

Marco, G.L., Petersen, N.S. & Stewart, E.E. (1983) *A test of the adequacy of curvilinear score equating models* in D. Weiss (ed.) *New horizons in testing: latent trait theory and computerized adaptive testing* New York: Academic Press

Markham, P.L. (1985) 'The rational deletion cloze and global comprehension in German' *Language Learning* 35.3: 423-430

Markosian, L.Z. and Ager, T.A. (1984) *Applications of parsing theory to computer-assisted instruction* in Wyatt (ed.)

Marshall, S.P. (1980) 'Procedural networks and production systems in adaptive diagnosis' *Instructional Science* 9: 129-143

Marxer, J.J. (1982) *Computer Storage and Retrieval of Test Items* in Curtis (ed.)

Masters, G.N. (1982) 'A Rasch model for partial credit scoring' *Psychometrika* 47: 149-174

Masters, G.N. (1984) 'Constructing an item bank using partial credit scoring' *Journal of Educational Measurement* 21: 19-32

Masters, G.N. and Wright, B.D. (1984) 'The essential process in a family of measurement models' *Psychometrika* 49.4: 529-544

McBride, J.R. and Weiss, D.J.A. (1974) 'A word knowledge item pool for adaptive ability measurement.' *Research Report* 74:2 Psychometric Methods Program, University of Minnesota

McDonald, R.P. (1981) 'The dimensionality of tests and items' *British Journal of Mathematical and Statistical Psychology* 34: 100-117

McIntyre, D. and Brown, S. (1978) 'The conceptualisation of attainment' *British Educational Research Journal* 4.2: 41-50

McIver, J.P. and Carmines, E.G. (1981) *Unidimensional Scaling* Beverley Hills: Sage Publications

McLeod, B. and McLaughlin, B. (1986) 'Restructuring or Automaticity? Reading in a

second language' *Language Learning* 36.2: 109–123

McNamara, D.R. (1979) 'Paradigm lost: Thomas Kuhn and educational research' *British Educational Research Journal* 5.2: 167–173

Meara, P. (1984) *Word Recognition in Foreign Languages* in Pugh and Ulijn (eds.)

Meijers, A.J.A. (1980) 'L'elaboration d'un cours de latin specialise pour des etudiants de theologie' *System* 8: 131–141

Mellenbergh, G.J. (1972) *Applicability of the Rasch model in two cultures* in L.J. Cronbach and P.J.D. Dreuth (eds.) *Mental Tests and Cultural Adaptation* The Hague: Mouton

Messick, S. (1975) 'The standard problem: meaning and values in educational measurement' *American Psychologist* 30: 955–966

Millman, J. (1974) *Test construction* in Lippey (ed)

Millman, J. (1980) *Computer-based Item Generation* in Berk (ed.)

Mitchell, D.C. (1982) *The Process of Reading* Chichester: John Wiley and Sons

Molenaar, I.W. (1981) 'On Wilcox's latent structure model for guessing' *British Journal of Mathematical and Statistical Psychology* 34: 224–228

Moller, A. (1985) *Project proposal for the development of an English placement test at Pusat Bahasa, USM* Kuala Lumpur: British Council mimeo

Morgan, J.L. (1982) *Discourse Theory and the Independence of Sentence Grammar* in Tannen (ed.)

Morgan, J.L. and Sellner, M.B. (1980) *Discourse and Linguistic Theory* in Spiro *et al.* (eds.)

Morrow, K.E. (1977) *Communicative Language Testing: Revolution or Evolution?* in Alderson and Hughes (eds.)

Moy, R.H. (1984) 'Proficiency standards and cut-scores for language proficiency' *System* 12.1: 13–24

Munby, J. (1978) *Communicative Syllabus Design: A Sociolinguistic Model for Defining the Content of Purpose-Specific Language Programmes* Cambridge: CUP

Neisser, U. (1967) *Cognitive Psychology* New York: The Century Psychology Series

Nelson, P. (1984) 'Towards a more communicative reading course: motivating students who are not "reading addicts"' *Reading in a Foreign Language* 2.1: 188-196

Newbould, C.A. and Massey, A.J. (1977) 'A computerized item-banking system (CIBS)' *British Journal of Educational Technology* 8.2: 114-123

New York State Department of Education (1975) *SPPED Cloze Exercises in a Multiple Choice Format* Prepared by the staff of the Bureau of School and Cultural Research, Divisions of Research and Evaluation: Albany, NY

Novick, M.R. (1983) *The centrality of Lord's Paradox and Exchangeability for all statistical inference* in Wainer and Messick (eds.)

Nunnally, J.C. (1978) *Psychometric Theory* (2nd Edition) New York: McGraw Hill

Nuttall, C. (1982) *Teaching Reading Skills in a Foreign Language* London: Heinemann

Odor, P. (1985) *CALL* Paper given to the Seminar on Educational Computing: Edinburgh University 18th April 1985

Oller, J.W. Jnr. (1972) 'Assessing competence in ESL reading' *TESOL Quarterly* 4.2: 107-116

Oller, J.W. Jnr. (1976) 'Evidence for a general proficiency factor: an expectancy grammar' *Die Neueren Sprachen* 2: 165-174

Oller, J.W. Jnr. (1979) *Language Tests at School* Longman

Oller, J.W. Jnr. and Perkins, K. (eds.) (1980) *Research in Language Testing* Rowley,Mass.: Newbury House

Olympia, P.L. Jnr. (1975) 'Computer-generation of truly repeatable examinations' *Educational Technology* 15.6: 53-55

O'Neil, H.F. Jnr. (ed.) (1979) *Procedures for Instructional Systems Development* New York: Academic Press

O'Reilly, R.P., Gorth, W.P. and Pinsky, P. (1973) 'CATC - an effort based on an evaluation methodology' *Educational Technology* 13.3: 32-34

Orleans, J.S. (1926) *A Study of the Nature of Difficulty* New York: Teachers College, Columbia University: Contributions to Education No. 206

Osburn, H.G. (1968) 'Item sampling for achievement testing' *Educational and Psychological Measurement* 28.1: 95–104

Osterlind, S.J. (1983) *Test Item Bias* Beverley Hills and London: Sage Publications

Paine, M. (1984) 'Vertical decalage and the EFL reader' *System* 12.1: 53–60

Palmer, A.S. and Bachman, L.F. (1981) *Basic Concerns in Test Validation* in Alderson and Hughes (eds.)

Palmer, F.R. (1974) *The English Verb* (2nd Edition) London: Longman

Pang, L.P. (1984) 'Is there a global factor of language proficiency? A critique of Oller and Hinofitis 1980' *IRAL* 22.3: 203–212

Parkinson, B., Mitchell, R. and Johnstone, R. (1982) *Mastery Learning in Foreign Language Teaching: a case study* Stirling Educational Monographs no. 8

Pask, G. (1976) *Conversation Theory: Applications in Education and Epistemology* Amsterdam: Elsevier

Patience, W. (1981) 'A comparison of latent trait models and equipercentile methods of vertically equating tests.' Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles.

Pearson, P.D. (ed.) (1984) *Handbook of Reading Research* New York: Longman

Pennock, C. (ed.) (1979) *Reading Comprehension at Four Linguistic Levels* Newark,Delaware: International Reading Association

Perfetti, C.A. (1983) *Reading Vocabulary and Writing: Implications for Computer-based Instruction* in Wilkinson (ed.)

Perkins, K. and Brutten, S.R. (1983) 'The effects of word frequency and contextual richness on ESL students' word identification abilities' *Journal of Research in Reading* 6.2: 119–128

Perkins, K. and Jones, B. (1985) 'Measuring passage contribution in ESL comprehension' *TESOL Quarterly* 19'1: 137–152

Perkins, K. and Pharis, K. (1980) *TOEFL scores in relation to standardised reading tests* in Oller and Perkins (eds.)

Petersen, N.S. and Novick, M.R. (1976) 'An evaluation of some models for culture-fair selection' *Journal of Educational Measurement* 13.1: 3-29

Piswanger, K. (1975) 'Interkulturelle Vergleiche mit dem Matrizentest von Formann' *Psychological Institute of Vienna*

Pollitt, A.B. (1979) *Item Banking* in *Issues in Educational Assessment* Scottish Education Department: HMSO 56-67

Pollitt, A.B. and Hutchinson, C. (1987) 'Calibrating graded assessments: Rasch partial credit analysis of performance in writing.' *Language Testing* 3: 72-92

Pollitt, A.B., Hutchinson, C., Entwistle, N. and De Luca, C. (1985) *What makes exam questions difficult?* Edinburgh: Scottish Academic Press

Popham, W.J. (1978) *Criterion-Referenced Measurement* Englewood Cliffs, New Jersey: Prentice Hall Inc.

Popham, W.J. (1980) *Domain Specification Strategies* in Berk (ed.)

Popyuk, W. (1980) 'A model for an item bank in second language proficiency testing' *System* 8: 47-52

Porter, D. (1983) *The effect of quantity of context on the ability to make linguistic predictions: a flaw in a measure of general proficiency* in Hughes and Porter (eds.)

Porter, D. (1983b) *Assessing communicative proficiency: the search for validity* in K. Johnson and D. Porter (eds.) *Perspectives in Communicative Language Teaching* London: Academic Press

Potthof, R.F. (1982) *Some issues in test equating* in P.W. Holland & D.B. Rubin (eds.) *Test equating* New York: Academic Press

Pratt, M.W., Krane, A.R. and Kendall, J.R. (1981) 'Triggering a schema: the role of italics and intonation in the interpretation of ambiguous discourse' *American Educational Research Journal* 18.3: 303-315

Prosser, F. (1974) *Item Banking* in Lippey (ed.)

Pugh, R.C. and Brunza, J.J. (1975) 'Effects of a confidence-weighted scoring system

on measures of test reliability and validity' *Educational and Psychological Measurement* 35: 73-78

Pugh, A.K. and Ulijn, J.M. (eds.) (1984) *Reading for Professional Purposes* London: Heinemann Educational

Pumfrey, P.D. (1976) *Reading: Tests and Assessment Techniques* London: Hodder and Stoughton

Purushothaman, M. (1975) *Secondary Mathematics Item Bank* Slough: NFER

Pusack, J.P. (1984) *Answer-processing and error correction in foreign language CAI* in Wyatt (ed.)

Pusack, J.P. and Otto, S.E.K. (1984) 'Blueprint for a comprehensive foreign language CAI currriculum' *Computers and the Humanities* 18: 195-204

Raatz, U. (1985) 'Better theory for better tests? *Language Testing* 2.1: 60-75

Raggett, M. St. J., Tutt, C., and Raggett, P. (eds.) (1979) *Assessment and Testing of Reading: Problems and Practices* London: Ward Lock Educational

Rasch, G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests* Expanded edition with a foreword and an afterword by B.D. Wright Chicago and London: The University of Chicago Press 1980

Rasch, G. (1966) 'An item analysis that takes individual differences into account' *British Journal of Mathematical and Statistical Psychology* 19: 49-57

Reckase, M.D. (1979) 'Unifactor latent trait models applied to multifactor tests: results and implications' *Journal of Educational Statistics* 4: 207-230

Ree, M.T. (1979) 'Estimating item characteristic curves' *Applied Psychological Measurement* 3: 371-385

Rentz, R.R. and Bashaw, W.L. (1975) *Equating Reading Tests with the Rasch model* (Vols I and II) Athens, Ga.: University of Georgia Educational Research Laboratory [ERIC documents nos. ED127330 and ED127331]

Rentz, R.R. and Bashaw, W.L. (1977) 'The national reference scale for reading: an application of the Rasch model' *Journal of Educational Measurement* 14: 161-179

Reynolds, R.E., Taylor, M.A., Steffenson, M.S., Shirey, L.L. and Anderson, R.C. (1982)

'Cultural schemata and reading comprehension' *Reading Research Quarterly* 17.3: 353-366

Richards, J.M. Jnr. (1967) 'Can computers write college admissions tests?' *Journal of Applied Psychology* 51.3: 211-215

Rindskopf, D. (1983) 'A general framework for using latent class analysis to test hierarchical and non-hierarchical learning models' *Psychometrika* 48: 85-97

Roid, G. (1979) *The Technology of Test-Item Writing* in O'Neil (ed.)

Roid, G. and Haladyna, T.M. (1978) 'A comparison of objective-based and modified-Bormuth item writing techniques' *Educational and Psychological Measurement* 38: 19-28

Roid, G.H. and Haladyna, T.M. (1982) *A Technology for Test-Item Writing* New York: Academic Press

Rosenblatt, L.M. (19780 *The Reader, the Text, the Poem: the Transactional Theory of the Literary Work* Carbondale,Illinois: Southern Illinois University Press

Rosenshine, B.V. (1980) *Skill Hierarchies in Reading Comprehension* in Spiro *et al.* (eds.)

Ross, J. (1966) 'An empirical study of a logistic mental test model' *Psychometrika* 31: 325-340

Rothkopf, E.Z. (1966) 'Learning from written instructive material: an exploration of the control of inspection behavior by test-like events' *American Educational Research Journal* 3: 241-249

Rothkopf, E.Z. and Kaplan, R. (1972) 'Exploring the effect of density and specificity of instructional objectives on learning from text' *Journal of Educational Psychology* 63: 295-302

Royer, J.M., Bates, J.A. and Konold, C.E. (1984) *Learning from text: methods of affecting reader intent* in Alderson and Urquhart (eds.)

Rummel, R.J. (1970) *Applied Factor Analysis* Evanston: Northwestern University Press

Ryan, E.B. (1981) *Identifying and remediating failures in reading comprehension:*

towards an instructional approach for poor comprehenders in Mackinnon and Waller (eds.)

Ryle, G. (1954) *Dilemmas* Cambridge: CUP

Salager, F. (1983) 'The lexis of fundamental medical English: classificatory framework and rhetorical function (a statistical approach)' *Reading in a Foreign Language* 1.1: 54-64

Salisnjak, J. (1973) 'Computer aided test preparation: six years of experience' *Educational Technology* 13.3: 37-38

Sanford, A.J. and Garrod, S.C. (1981) *Understanding Written Language* Chichester: John Wiley and Sons

Scandura, J.M. (1977) *Problem Solving* London: Academic Press

Self, J.A. (1979) 'Student models and artificial intelligence' *Computers and Education* 3: 309-312

Seliger, H.W. (1985) 'Testing authentic language: the problem of meaning' *Language Testing* 2.1: 1-15

Shankweiler, D. and Liberman, I.Y. (1972) *Misreading: A Search for Causes* in Kavanagh and Mattingley (eds.)

Shaw, K.E. (1976) 'Paradigms or contested concepts?' *British Journal of Educational Technology* 7.2: 18-24

Shoemaker, D.M. (1975) 'Toward a framework for achievement testing' *Review of Educational Research* 45: 127-147

Shoemaker, D.M. (1976) *Applicability of item banking and matrix sampling to educational assessment* in de Gruijter and van der Kamp (eds.)

Shohamy, E. (1984) 'Does the testing method make a difference? The case of reading comprehension' *Language Testing* 1.2: 147-170

Sim, D. and Bensoussan, M. (1979) *Control of conceptualized function and content words as it affects EFL reading comprehension test scores* in Mackay *et al.* (eds.)

Singer, H. and Ruddell, R.B. (eds.) (1976) *Theoretical Models and Processes of Reading* Newark,Delaware: International Reading Association

Skaggs, G. and Lissitz, R.W. (1986) 'IRT test equating: relevant issues and a review of recent research' *Review of Educational Technology* 56.4: 495–529

Skehan, P. (1983) *Review* of Johnston, P.H. (1983) *Reading in a Foreign Language* 1.2: 133–136

Skehan, P. (1984) 'Issues in the testing of English for specific purposes' *Language Testing* 1.2: 202–220

Slade, P.D. and Dewey, M.E. (1983) 'Role of grammatical clues in multiple choice questions: an empirical study' *Medical Teacher* 5.4: 146–148

Slinde, J.A. and Linn, R.L. (1977) 'Vertically equated tests: fact or phantom?' *Journal of Educational Measurement* 14: 23–32

Slinde, J.A. and Linn, R.L. (1978) 'An exploration of the adequacy of the Rasch model for the problem of vertical equating' *Journal of Educational Measurement* 15: 23–35

Slinde, J.A. and Linn, R.L. (1979) 'A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty' *Journal of Educational Measurement* 16: 159–165

Slocum, T.J. (1972) *Locally Produced Item Banks* in Curtis (ed.)

Smith, F. (1971) *Understanding Reading* New York: Holt, Rinehart and Winston

Smith, J.M., Smith, D.E.P. and Brink, J.R. (1977) *A technology of reading and writing. Vol.2: Criterion-referenced tests for reading and writing* New York; Academic Press

Smith, R.M. (1985) 'A comparison of Rasch person analysis and robust estimators' *Educational and Psychological Measurement* 45.3: 433–444

Spache, G. (1953) 'A new readability formula for primary grade reading materials' *Elementary School Journal* 53: 410–413

Spearritt, D. (1972) 'Identification of subskills of reading comprehension by maximum likelihood factor analysis' *Reading Research Quarterly* 8: 92–111

Spiro, R.J. (1980) *Constructive Processes in Prose Comprehension and Recall* in Spiro *et al.* (eds.)

Spiro, R.J., Bruce, B.C. and Brewer, W.F. (eds.) (1980) *Theoretical Issues in Reading*

Comprehension Hillsdale, New Jersey: Lawrence Erlbaum Associates

Spolsky, B. (1981) *Some Ethical Questions About Language Testing* in Klein-Braley and Stevenson (eds.)

Spolsky, B. (1985) 'The limits of authenticity in language testing' *Language Testing* 2.1: 31-40

Spolsky, B. (1985) 'What does it mean to know how to use a language? An essay on the theoretical basis of language testing' *Language Testing* 2.2: 180-191

Steadman, S. and Gipps, C. (1984) 'Teachers and testing: pluses and minuses' *Educational Research* 26.2: 121-126

Steffensen, M. (1986) 'Register, cohesion, and cross-cultural reading comprehension' *Applied Linguistics* 7.1: 71-85

Steffenson, M.S. and Joag-Dev, C. (1984) *Cultural Knowledge and Reading* in Alderson and Urquhart (eds.)

Stern, H.H. (1983) *Fundamental Concepts in Language Teaching* Oxford: OUP

Sternberg, R.J., Powell, J.S. and Kaye, D.B. (1983) *Teaching Vocabulary - Building Skills* in Wilkinson (ed.)

Stevenson, D.K. (1982) *"All of the above": on problems in the testing of FL reading* in Jung (ed.)

Sticht, T.G. (1972) *Learning by Listening* in Freedle and Carroll (eds.)

Stodola, Q.C. (1973) 'Use of computer-assembled tests in the California State University and College system' *Educational Technology* 13.3: 40-41

Stodola, Q.C. (1974) *Item Classification and Selection* in Lippey (ed.)

Stokes, A. (1978) *The Reliability of Readability Formulae* in A.K. Pugh, V.J. Lee and J. Swann (eds.) *Language and Language Use* London: Heinemann Educational

Sumner, R. (1975) '"Mastery learning": an all or nothing?' *British Educational Research Journal* 1.2: 24-25

Suppes, P. (ed.) (1981) *Computer-Assisted Instruction at Stanford: 1968-1980* Stanford University: Institute for Mathematical Studies in the Social Sciences

Swain, M., Lapkin, S. and Barik, H.C. (1976) 'The cloze test as a measure of second language proficiency for young children' *Working Papers on Bilingualism* 11: 32–42

Swan, M. (1985) 'A critical look at the communicative approach (1)' *ELTJ* 39.1: 2–12

Swan, M. (1985b) 'A critical look at the communicative approach (2)' *ELTJ* 39.2: 76–87

Tall, G. (1981) *The possible dangers of applying the Rasch model to school examinations and standardized tests* in Lacey and Lawton (eds.)

Tannen, D. (ed.) (1982) *Analyzing Discourse: Text and Talk* Washington D.C.: Georgetown University Press

Tansley, A.E. (1967) *Reading and Remedial Reading* London: Routledge and Kegan Paul

Theunissen, T.J.J.M. (1987) 'Text banking and test design' *Language Testing* 3: 1–8

Thimbleby, H. (1979) 'Computers and Human Consciousness' *Computers and Education* 3: 241–243

Thornkike, E.L. (1917) 'Reading as reasoning: a study of mistakes in paragraph reading' *Journal of Educational Psychology* 8: 323–332

Thorndike, R.L. (1973) *Reading Comprehension Education in 15 Countries: an Empirical Study* International Studies in Evaluation III New York: John Wiley and Sons

Thorndike, R.L. (1974) 'Reading as reasoning' *Reading Research Quarterly* 9.2: 135–147

Thorndike, R.L. (1982) *Applied Psychometrics* Boston: Houghton Mifflin Company

Thornkike, R.L. (1983) *How can we practice what we preach?* in Wainer and Messick (eds.)

Thurstone, L.L. (1956) 'Note on a reanalysis of Davis' reading tests' *Psychometrika* 11: 185–188

Tinsley, H.E. and Dawis, R.V. (1975) 'An investigation of the Rasch simple logistic model: sample-free item and test calibration.' *Educational and Psychological Measurement* 35: 325– 339

Tobin, Y. (1984) 'Applying two linguistic theories to improve reading comprehension in German' *IRAL* 22.2: 95–108

Toggenburger, F. (1973) 'Classroom teacher support system' *Educational Technology* 13.3: 42–43

Trabasso, T. (1980) *On the making of inferences during reading and their assessment* Urbana, Illinois: Center for the Study of Reading (Tech. Rep. No. 157) [ED 181429]

Tuinman, J.J. (1974) 'Determining the passage dependency of comprehension questions in 5 major tests' *Reading Research Quarterly* 9.2: 206–223

Tuinman, J.J. (1979) 'Beyond criterion-referenced measurement' in M.L. Kamil and A.J. Moe (eds.) *Reading Research: Studies and Applications* Clemson, S. Carolina: National Reading Conference Inc.

Tumposky, N.R. (1984) 'Behavioural objectives, the cult of efficiency and foreign language learning: are they compatible?' *TESOL Quarterly* 18.2: 295–310

Ulijn, J. (1980) 'Foreign language reading research: recent trends and future prospects' *Journal of Research in Reading* 3.1: 17–37

Ulijn, J.M. (1984a) 'Foreign language (FL) reading: conceptual and syntactic strategies and their consequences for the role of the native language (NL)' *IRAL* 22.1: 71–73

Ulijn, J.M. (1984b) *Reading for professional purposes: psycholinguistic evidence in a cross-linguistic perspective* in Pugh and Ulijn (eds.)

Ulijn, J.M. (1985) *Review* of Nuttall (1982) *Applied Linguistics* 6.1: 91–92

Urry, V.W. (1977) 'Tailored testing: a successful application of latent trait theory' *Journal of Educational Measurement* 14.2: 181–196

Vale, C.D. and Weiss, D.J. (1975) *A study of computer administered stradaptive ability testing* Res. Rep. 75-4 Psychometric Methods Program, University of Minnesota

Valette, R.M. (1977) *Modern Language Testing* 2nd Edition New York: Harcourt, Brace, Jovanovich

van der Kamp, L.J.Th. (1976) *Generalizability and Educational Measurement* in de

Gruijter and van der Kamp (eds.)

Van Parreren, C.F. and Schouten-Van Parreren, M.C. (1982) *Contextual Guessing: A Trainable Reader Strategy* in Jung (ed.)

Venezky, R.L. (1983) *Evaluating Computer-Assisted Instruction on its Own Terms* in Wilkinson (ed.)

Venezky, R.L., Calfee, R.C. and Chapman, R.S. (1970) *Skills Required for Learning to Read* in Gunderson (ed.)

Vickers, F.D. (1973) 'Generative test generators' *Educational Technology* 13.3: 43-44

Vollmer, H.J. (1979) *Why are we interested in general language proficiency?* in Alderson and Hughes (eds.)

Vollmer, H.J. (1981) *Issue or non-issue: general language proficiency revisited* in Alderson and Hughes (eds.)

Vollmer, H.J. (1983) *The Structure of Foreign Language Competence* in Hughes and Porter (eds.)

Wainer, H. and Messick, S. (eds.) (1983) *Principals [sic] of Modern Psychological Measurement* Hillsdale, New Jersey: Lawrence Erlbaum Associates

Walsh, V. (1980) 'Reading scientific texts in English' *System* 8: 231-239

Walton, L., Harlow, L., Smith, W.F., Boyle, T.A. and Walker, J. (1979) 'An evaluation and placement technique: the branched program achievement test' *System* 7: 211-217

Wanat, S.F. (ed.) (1977a) *Issues in Evaluating Reading* [Papers in Applied Linguistics, Linguistics and Reading Series 1] Arlington, Va.: Center for Applied Linguistics

Wanat, S.F. (ed.) (1977b) *Language and Reading Comprehension* [Papers in Applied Linguistics, Linguistics and Reading Series 2] Arlington, Va.: Center for Applied Linguistics

Wardough, R. (1969) *Reading: A Linguistic Perspective* New York: Harcourt, Brace and World Inc.

Watson, D.M. (1984) 'Computer-assisted learning for school pupils of History, French and English in the U.K.' *Computers and the Humanities* 18: 233-241

Webber, B.L. (1980) *Syntax Beyond the Sentence: Anaphora* in Spiro *et al.* (eds.)

Weir, C.J. (1983) 'Identifying the language problems of overseas students in tertiary education in the U.K.' Unpublished Ph.D. thesis: University of London

Weiss, D.J. (1979) *Computerized Adaptive Achievement Testing* in O'Neil (ed.)

Wellens, B. (1972) *Publisher's Role in Preparation of Items* in Curtis (ed.)

Whitely, S.E. (1977) 'Models, meanings and misunderstandings: some issues in applying Rasch's theory' *Journal of Educational Measurement* 14: 227 - 235

Whitely, S. and Dawis, R. (1974) 'The nature of objectivity with the Rasch model' *Journal of Educational Measurement* 2.2: 163-178

Widdowson, H.G. (1978) *Teaching Language As Communication* Oxford: OUP

Widdowson, H.G. (1984) *Reading and Communication* in Alderson and Urquhart (eds.)

Wiener, M. and Cromer, W. (1970) *Reading and Reading Difficulty: a conceptual analysis* in Gunderson (ed.)

Wilkinson, A.C. (ed.) (1983) *Classroom Computers And Cognitive Science* New York: Academic Press

Wilkinson, A.C. and Patterson, J. (1983) *Issues at the interface of theory and practice* in Wilkinson (ed.)

Williams, E. (1985) *Review* of H.S. Marden (1983) *Techniques in testing* New York and Oxford: OUP *Language Testing* 2.1: 105-111

Williams, R. (1983) 'Teaching the recognition of cohesive ties in reading in a foreign language' *Reading in a Foreign Language* 1.1: 35-52

Willmott, A.S. (1976) 'The place of item banks in local research' *British Educational Research Journal* 2.2: 40-42

Willmott, A.S. and Fowles, D.E. (1974) *The Objective Interpretation of Test Performance: the Rasch model applied* Slough: National Foundation for Educational Research

Wilson, J.B. (1972) *Philosophy and Educational Research* Slough: National

Foundation for Educational Research

Wood, R. (1973) 'Resonse-contingent testing' *Review of Educational Research* 43: 529-544

Wood, R. (1976) *Trait Measurement and Item Banks* in de Gruijter and van der Kamp (eds.)

Wood, R. (1978) 'Fitting the Rasch model - a heady tale' *British Journal of Mathematical and Statistical Psychology* 31: 27-32

Wood, R. and Skurnik, L.S. (eds.) (1969) *Item Banking* Slough: National Foundation for Educational Research

Woods, A. *Principal components and factor analysis in the investigation of the structure of language proficiency* in Hughes and Porter (eds.)

Woods, A. and Baker, R. (1985) 'Item response theory' *Language Testing* 2.2: 119-140

Wright, B.D. (1968) 'Sample-free test calibration and person measurement' *Proceedings of the 1967 Invitational Conference on Testing Problems* 85-101

Wright, B.D. (1977a) 'Misunderstanding the Rasch model' *Journal of Educational Measurement* 14: 219-225

Wright, B.D. (1977b) 'Solving measurement problems with the Rasch model' *Journal of Educational Measurement* 14.2: 97-116

Wright, B.D. and Bell, S.R. (1981) *Fair and useful testing with item banks* Research Memorandum 32, Chicago: University of Chicago Department of Education, MESA Psychometrics Laboratory

Wright, B.D. and Douglas, G.A. (1977) 'Conditional versus unconditional procedures for sample-free item analysis' *Educational and Psychological Measurerement* 37: 573-586

Wright, B.D. and Masters, G.N. (1982) *Rating Scale Analysis* Chicago: MESA Press

Wright, B.D. and Mead, R.J. (1977) *BICAL: Calibrating items and scales with the Rasch model* Research Memorandum 23, Statistical Laboratory, Department of Education, University of Chicago

Wright, B.D., Mead, R.J. and Bell, S.R. (1980) *BICAL: Calibrating items with the Rasch model* Research Memorandum 23c, Statistical Laboratory, Department of Education, University of Chicago

Wright, B.D. and Panchapakesan, N. (1969) 'A procedure for sample-free item analysis' *Educational and Psychological Measurement* 29: 23-48

Wright, B.D. and Stone, M.H. (1979) *Best Test Design* Chicago: MESA Press

Wu, P.E.-Shi (1981) *Construction and evaluation of a computer-assisted instruction curriculum in spoken Mandarin* in Suppes (ed.)

Wyatt, D.H. (ed.) (1984) *Computer-Assisted Language Instruction* Oxford: Pegamon

Yorio, C.A. (1971) 'Some sources of reading problems for foreign language learners' *Language Learning* 21.1: 107-115

Zeller, R.A. and Carmines, E.G. (1980) *Measurement in the Social Sciences* Cambridge: Cambridge University Press

Zuck, L.V. and Zuck, J.G. (1984) *The Main Idea: specialist and non- specialist judgments* in Pugh and Ulijn (eds.)

Appendix I
Test Content

I. Test Content

The content description for each item in all versions of the test is as follows [with item group label as used in factor analysis in Chapter 7]: Part 1

1. Past/Perfect Verb Forms[SUBSC 1]

    a. Form A: 8 - 11 - 18 - 23 - 29

    b. Form B: 3 - 37 - 40 - 49 - 50

    c. Form C: 7 - 14 - 18 - 45 - 48

    d. Form D: 30 - 37 - 41 - 47 - 48

2. Present/Future Verb Forms [SUBSC 2]

    a. Form A: 2 - 9 - 30 - 39 - 44

    b. Form B: 4 - 7 - 20 - 22 - 39

    c. Form C: 5 - 35 - 36 - 41 - 46

    d. Form D: 7 - 17 - 40 - 42 - 50

3. Conditionals [SUBSC 3]

    a. Form A: 1 - 4 - 5 - 26 - 48

    b. Form B: 2 - 6 - 8 - 16 - 48

    c. Form C: 6 - 8 - 43 - 44 - 47

    d. Form D: 3 - 12 - 21 - 35 - 43

4. Prepositions [SUBSC 4]

    a. Form A: 6 - 45 - 46 - 47 - 49

    b. Form B: 1 - 19 - 28 - 29 - 43

    c. Form C: 16 - 24 - 31 - 34 - 40

    d. Form D: 9 - 20 - 36 - 45 - 49

5. Conjunctions [SUBSC 5]

a. Form A: 3 - 7 - 15 - 16 - 28

b. Form B: 18 - 32 - 42 - 44 - 46

c. Form C: 11 - 22 - 29 - 32 - 37

d. Form D: 2 - 14 - 19 - 24 - 39

6. Relative Pronouns [SUBSC 6]

a. Form A: 21 - 34 - 35 - 41 - 50

b. Form B: 15 - 25 - 27 - 30 - 45

c. Form C: 1 - 10 - 21 - 25 - 27

d. Form D: 6 - 25 - 27 - 34 - 38

7. Articles [SUBSC 7]

a. Form A: 10 - 19 - 24 - 25 - 37

b. Form B: 5 - 11 - 17 - 24 - 33

c. Form C: 2 - 23 - 26 - 30 - 42

d. Form D: 10 - 15 - 18 - 28 - 44

8. Modals/Auxiliaries [SUBSC 8]

a. Form A: 13 - 22 - 27 - 42 - 43

b. Form B: 9 - 10 - 13 - 14 - 23

c. Form C: 4 - 13 - 15 - 19 - 38

d. Form D: 4 - 5 - 26 - 32 - 46

9. Infinitives/Gerunds [SUBSC 9]

a. Form A: 31 - 32 - 36 - 38 - 40

b. Form B: 21 - 26 - 34 - 36 - 47

c. Form C: 3 - 12 - 20 - 39 - 50

d. Form D: 1 - 8 - 11 - 23 - 31

10. Comparisons [SUBSC 10]

a. Form A: 12 - 14 - 17 - 20 - 33

b. Form B: 12 - 31 - 35 - 38 - 41

c. Form C: 9 - 17 - 33 - 28 - 49

d. Form D: 13 - 16 - 22 - 29 - 33

Part 2

1. Active/Passive [INT 1]

a. Form A: 62 - 69 - 70

b. Form B: 57 - 69 - 72

c. Form C: 58 - 66 - 67

d. Form D: 51 - 60 - 71

2. Cause/Effect [INT 2]

a. Form A: 51 - 53 - 73

b. Form B: 68 - 70 - 73

c. Form C: 52 - 63 - 65

d. Form D: 52 - 66 - 73

3. Purpose/Result [INT 3]

a. Form A: 58 - 71 - 72

b. Form B: 54 - 64 - 66

c. Form C: 60 - 68 - 64

d. Form D: 58 - 61 - 72

4. Reported Speech [INT 4]

a. Form A: 54 - 60 - 67

b. Form B: 52 - 60 - 67

c. Form C: 56 - 69 - 72

d. Form D: 57 - 69 - 70

5. Similarity/Difference [INT 5]

a. Form A: 56 - 66

b. Form B: 59 - 65

c. Form C: 54 - 70

d. Form D: 54 - 63

6. Sequence of Events [INT 6]

a. Form A: 61 - 68

b. Form B: 51 - 58

c. Form C: 53 - 62

d. Form D: 56 - 65

7. Relative Clauses [INT 7]

a. Form A: 55 - 65

b. Form B: 61 - 62 - 63

c. Form C: 59 - 61 - 71

d. Form D: 55 - 67

8. Connectors [INT 8]

    a. Form A: 52 - 57

    b. Form B: 55 - 71

    c. Form C: 55 - 57

    d. Form D: 62 - 68

9. Possibility/Certainty [INT 9]

    a. Form A: 59 - 63 - 64

    b. Form B: 53 - 56

    c. Form C: 51 - 73

    d. Form D: 53 - 59 - 64

10. Text Types [INT 10]: items 74 to 77 inclusive on all forms

11. Logical Ordering [INT 11]: items 78 to 84 inclusive on all forms

12. Information Transfer [INT 12]: items 85 to 94 inclusive on all forms

13. Connectors in Discourse [INT 13]: items 95 to 100 inclusive on all forms

Part 3

1. Recognising Sequence [COMP 1]

    a. Form A: none

    b. Form B: none

    c. Form C: 103 - 136

    d. Form D: none

2. Recognising Words in Context [COMP 2]

a. Form A: 106 - 107 - 114 - 131 - 138 - 139

b. Form B: 106 - 115 - 122 - 130 - 131 - 138 - 139

c. Form C: 106 - 108 - 114 - 115 - 122 - 138

d. Form D: 107 - 114 - 122 - 132 - 138 - 139

3. Identifying the Main Idea [COMP 3]

a. Form A: 101 - 102 - 105 - 109 - 113 - 130 - 136 - 137

b. Form B: 101 - 105 - 109 - 112 - 114 - 120 - 121 - 125 - 129 - 134 - 137

c. Form C: 109 - 110 - 113 - 117 - 121 - 133 - 137

d. Form D: 101 - 104 - 106 - 117 - 121 - 126 - 129 - 134 - 137

4. Decoding Detail [COMP 4]

a. Form A: 103 - 112 - 125

b. Form B: 104 - 110 - 119 - 136

c. Form C: 101 - 102 - 111 - 120 - 135

d. Form D: 102 - 109 - 111 - 112 - 120 - 125 - 136

5. Drawing Inferences [COMP 5]

a. Form A: 104 - 108 - 111 - 115 - 116 - 123 - 124- 126 - 132 - 133 - 140

b. Form B: 103 - 107 - 108 - 111 - 116 - 118 - 123 - 124 - 128 - 132 - 133 - 135 - 140

c. Form C: 104 - 107 - 112 - 116 - 119 - 123 - 124 - 131 - 132 - 134 - 139 - 140

d. Form D: 103 - 108 - 115 - 116 - 119 - 123 - 124 - 130 - 131 - 133 - 135 - 140

6. Recognising Cause and Effect [COMP 6]

a. Form A: 117 - 121 - 122 - 129 - 135

b. Form B: 113 - 117

c. Form C: 105 - 125 - 129 - 130

d. Form D: 105 - 110 - 113 - 127 - 128

7. Comparing and Contrasting [COMP 7]

a. Form A: 110 - 118 - 119 - 120 - 127 - 128 - 134

b. Form B: 102 - 126 - 127

c. Form C: 118 - 126 - 127 - 128

d. Form D: 118

Source of texts for Part 3

1. Form A

a. Dubin and Olshtain pp. 177 - 178

b. Mosback and Mosback pp. 81 - 82

c. Heaton pp. 35 - 36

d. Dubin and Olshtain pp. 256 - 257

e. Reading and Thinking in English 4 pp. 87 - 88

2. Form B

a. Mosback and Mosback pp. 77 - 78

b. Dubin and Olshtain pp. 206 - 207

c. Heaton pp. 144 - 146

d. Young and Gardner pp. 46 - 47

e. Reading and Thinking in English 4 pp. 59 - 60

3. Form C

a. Heaton pp. 220 - 221

b. Dubin and Olshtain pp. 177 - 178

c. Mosback and Mosback pp. 81 - 82

d. Heaton pp. 35 - 36

e. Dubin and Olshtain pp. 141 - 142

4. Form D

a. Dubin and Olshtain pp. 206 - 207

b. Heaton pp. 234 - 235

c. Mosback and Mosback pp. 77 - 78

d. Reading and Thinking in English 4 p. 42

e. Reading and Thinking in English 4 pp. 59 -60

References British Council (1980) *Reading and Thinking in English Book 4: Discourse in Action* Oxford: OUPP

Dubin, F. and Olshtain, E. (1981) *Reading by all means* Reading, Mass.: Wesley Publishing Company

Ewer, J.R. and Latorre, G. (1969) *A course in basic scientific English* London: Longman

Heaton, J.B. (1984) *Create and Communicate: Book 4 (Revised Edition)* Singapore: Longman

Appendix II
Test Forms C and D
plus additional Form A and B Part 3 items

UNIVERSITI SAINS MALAYSIA
PUSAT BAHASA DAN TERJEMAHAN

ENGLISH LANGUAGE PLACEMENT TEST
FORM C

PART 1 - (25 minutes)

Part 1 is a test of your knowledge of basic English grammar. There are 50 questions.

For each question choose the answer - (a), (b), (c) or (d) - which best completes the sentence.

Mark your answer in thick pencil on the answer sheet.

For example:

Question 104.

This camera_____ good pictures.

(a) sees
(b) makes
(c) takes
(d) looks at

The correct answer is (c) - This camera takes good pictures.

You mark your answer sheet like this

104 

Now turn over and begin Part 1. Work as quickly as possible.

.../-

1. He wanted to come home at 2 o'clock, _____ didn't suit me at all.
   (a) who (b) which
   (c) that (d) of which

2. Those who study _____ law can expect to do a lot of reading.
   (a) the (c) any
   (b) a (d) ____

3. Can we ever succeed _____ nuclear weapons?
   (a) to abolish (c) in abolishing
   (b) by abolishing (d) and abolish

4. This play _____ written by Shakespeare, because the style is Milton's.
   (a) was (b) couldn't have been
   (c) could have been (d) need have been

5. Don't leave the aeroplane, until the steward _____ you to.
   (a) will tell (b) would tell
   (c) is telling (d) tells

6. "If you need emergency treatment _____ to the clinic".
   (a) you will go (c) you would go
   (b) you are going (d) go

7. Last year he _____ to the United States.
   (a) has been (c) went
   (b) has gone (d) was

8. Unless the floods _____ we shan't be able to use the ford.
   (a) will go down (c) have gone down
   (b) had gone down (d) are going down.

9. _____ what was budgeted, we actually spent very little.
   (a) Similar to (c) Unlike
   (b) In contrast to (d) In the same way as

10. The roads were crowded with refugees, _____ many were injured.
   (a) who (b) which
   (c) by which (d) of whom

11. We can still go and see him _____ it is raining.
   (a) even though (b) whereas
   (c) besides (d) despite

...3.

12. The captain was the last man _____ the sinking ship.
(a) leaving (b) left
(c) to leave (d) by leaving

13. This creature has 8 legs so it _____ an insect.
(a) mustn't be (c) should have been
(b) needn't be (d) can't be

14. When he _____ his wife off at the airport, he went back home.
(a) saw (b) has seen
(c) was seeing (d) had seen

15. I'm sorry the essay is so bad; I _____ half asleep when I wrote it
(a) must have been (b) could have been
(c) must be (d) need to be

16. He is fully aware _____ the difficulties in this situation.
(a) of (b) by
(c) on (d) to

17. Social insects live in integrated communities which in some ways _____ human communities
(a) similarly (b) are similar to
(c) differ from (d) by comparison

18. When I first _____ to this area it was a very quiet place.
(a) have come (b) had come
(c) came (d) was coming

19. The result of your experiment is very unusual - you _____ made a mistake.
(a) should have (c) need to have
(b) must have (d) can have

20. The police were unable to prevent the robbery _____ place.
(a) to take (b) from take
(c) taken (d) taking

21. Chlorine is a green gas _____ dissolves readily in water.
(a) who (c) what
(b) which (d) ____

22. The course of medicine is not the easiest; _____ it is the most difficult.
(a) on the other hand (c) on the contrary
(b) and yet (d) in addition

...4/.

23. Before you visit France you should learn _____ language.
(a) a (c) the
(b) some (d) _____

24. After they arrived _____ the airport they found the plane had been delayed.
(a) to (c) at
(b) in (d) inside

25. There wasn't directory in the telephone box _____ I was calling.
(a) which (b) at which
(c) from which (d) whom

26. _____ history of Australia is often said to be based on lies.
(a) The (b) A
(c) Some (d) Any

27. The boy _____ John shared a flat with was a philosophy student.
(a) which (c) with whom
(b) whose (d) _____

28. _____ animals, plants do not move.
(a) Like (b) As
(c) By comparison (d) Unlike

29. He is still fit _____ he is not so young.
(a) despite (c) even though
(b) whereas (d) while

30. If you had _____ sense you wouldn't leave your car unlocked.
(a) a (c) any
(b) some (d) the

31. Turn the key _____ the lock until it clicks.
(a) within (c) at
(b) into (d) in

32. _____ his age he is still very active.
(a) Although (b) Nevertheless
(c) In spite of (d) Besides

...5/-

33. Aluminium is light _____ copper is not so light.
(a) whereas (c) furthermore
(b) indeed (d) thus

34. He tried to drive away the wolves by throwing stones _____ them.
(a) to (b) in
(c) for (d) at

35. I can't let you drive the car until you _____ a licence.
(a) get (b) will get
(c) would get (d) are getting

36. Heat the oil until it _____ to bubble.
(a) began (c) will begin
(b) begins (d) would begin

37. I'll stay here _____ I get an answer.
(a) by the time that (b) until
(c) despite (d) although

38. A poem _____ sonnet unless it has 14 lines
(a) doesn't need to be (b) can't be
(c) mustn't be (d) needs to be

39. I saw him _____ the road and enter the shop.
(a) cross (c) to cross
(b) crossing (d) had crossed

40. He is envious _____ those who drive fast cars.
(a) for (c) with
(b) of (d) by

41. By the end of the semester, I _____ all twelve volumes.
(a) read (b) would read
(c) will have read (d) will read

42. The study of _____ literature is more than just reading stories.
(a) a (b) the
(c) some (d) ____

43. If he gets a work permit, he _____ for another 6 months.
(a) stays (b) will stay
(c) would stay (d) would have stayed

44. If he _____ more exercise he wouldn't be so unhealthy.
(a) took (b) would take
(c) will take (d) takes

45. Who _____ last year?
(a) teaches (c) had taught
(b) has taught (d) taught

46. By the end of next week _____ my preliminary training.
(a) I'll have finished (c) I am finish
(b) I finish (d) I'll finish

47. The disease will spread unless you _____ the carriers.
(a) isolate (b) would isolate
(c) will isolate (d) would have isolated

48. I _____ breakfast when the phone rang.
(a) had (c) have had
(b) was having (d) have

49. Science is subjective within man's nature _____ humanistic studies are.
(a) so as (c) by comparison
(b) similarly (d) in the same way that

50. I _____ a lot but I don't have time now.
(a) used to ride (c) am used to riding
(b) used to riding (d) am used to ride

UNIVERSITI SAINS MALAYSIA
PUSAT BAHASA DAN TERJEMAHAN

ENGLISH LANGUAGE PLACEMENT TEST
FORM C

PART 2 - (35 minutes)

Part 2 is a test of your working and reading grammar of English. There are 50 questions.

Not all the questions are of the same type, but you should in every case mark your answer - (a), (b). (c), or (d) - in thick pencil on the answer sheet.

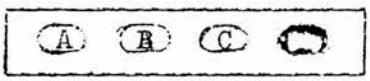
For example:

Which sentence is closest in meaning to the first sentence?

105. John likes neither tea or coffee.

(a) John likes both tea and coffee.
(b) John likes tea but not coffee
(c) John likes coffee but not tea
(d) John doesn't like coffee or tea.

The correct answer is (d) - John doesn't like coffee or tea. You mark your answer sheet like this:

105 

Now turn over and begin Part 2. Work as quickly as possible.

51. It is envisaged that man will soon make contact with extra-terrestrial life

(a) People predict that man will soon make contact with extra-terrestrial life.

(b) People hope that man will soon make contact with extra-terrestrial life.

(c) We know that we will soon make contact with extra-terrestrial life.

(d) It is expected that we will soon make contact with extra-terrestrial life.

52. The fire isn't very hot. It won't boil a kettle.

(a) The fire is too hot to boil a kettle.

(b) The fire isn't hot enough to boil a kettle.

(c) The fire is so hot it won't boil a kettle.

(d) The fire isn't cool enough to boil a kettle.

53. Earlier, he had found a small piece of broken pottery.

(a) He had found the pottery in the morning.

(b) He had found the pottery before it was too late.

(c) He had found the pottery at a previous moment.

(d) He had found the pottery just in time.

54. Unlike the desert annuals, the perennials have special features which enable them to survive as plants for several years.

(a) Desert annuals can survive for several years.

(b) Perennials do not live in the desert.

(c) Perennials have special features which annuals do not.

(d) Annuals are similar to perennials in that they both live in the desert.

55. Tomatoes, besides having a rich flavour, contain sugar and vitamins.

(a) As well as they have a rich flavour, tomatoes contain sugar and vitamins.

(b) Apart from they contain a rich flavour, tomatoes contain sugar and vitamins.

(c) Tomatoes having a rich flavour contain sugar and vitamins.

(d) In addition to having a rich flavour, tomatoes contain sugar and vitamins.

56. "Let's go to a cinema," said Ann.

(a) Ann advised us to go to a cinema.

(b) Ann suggested that we go to a cinema.

(c) Ann told us to go to a cinema.

(d) Ann recommended that we go to a cinema.

...3/-

57. Walk very carefully over the floor. Otherwise you may fall.

(a) If you don't walk carefully you won't fall.

(b) Unless you walk carefully, don't fall.

(c) Walk very carefully unless you fall.

(d) Walk very carefully lest you fall.

58. Somebody had cleaned my shoes and brushed my suit.

(a) My shoes and suit had been cleaned by somebody and than brushed.

(b) My shoes had been cleaned by somebody but I brushed my own suit.

(c) I had my shoes cleaned any my suit brushed.

(d) My shoes had been cleaned and my suit brushed.

59. The policeman caught the criminal but was injured in the fight.

(a) The criminal who was injured in the fight was caught by the policeman.

(b) The policeman, who was injured in the fight, caught the criminal.

(c) The policeman who was injured in the fight caught the criminal.

(d) The criminal, who was injured in the fight, was caught by the policeman.

60. Seat belt laws were introduced to help prevent accidents.

(a) Accidents have been prevented by the introduction of seat belt laws.

(b) Seat belt laws have led to the prevention of accidents.

(c) Seat belt laws were introduced so that accidents might be prevented.

(d) Because seat belt laws were introduced, accidents have been prevented.

61. He didn't thank us, which offended us.

(a) He must have thanked us.

(b) He needn't have thanked us.

(c) He should have thanked us.

(d) He didn't need to thank us.

62. The soaked lentils should be simmered in a large saucepan.

(a) Having simmered the lentils, then soak them.

(b) Soak the lentils at the same time as you simmer them.

(c) Simmer the lentils; next, soak them.

(d) Soak the lentils then simmer them.

.../-

63. He works too fast; that's why he makes so many mistakes.

(a) If he didn't work so fast he wouldn't make so many mistakes.

(b) If he worked faster he wouldn't make so many mistakes.

(c) If he works too fast, he'll make so many mistakes.

64. If she had paid the fine she wouldn't have been sent to prison.

(a) She paid the fine and was sent to prison.

(b) She paid the fine but was sent to prison.

(c) She didn't pay the fine and wasn't sent to prison.

(d) She didn't pay the fine and was sent to prison.

65. The prices in hotels have risen alarmingly; hence tourists are now going elsewhere.

(a) Because of the alarming rise in hotel prices tourists are now going elsewhere.

(b) So that tourists will now go elsewhere, the hotels have raised their prices alarmingly.

(c) Tourists who have found alarming increases in hotel prices are now going elsewhere.

(d) Because tourists are now going elsewhere, hotels have raised their prices alarmingly.

66. The judge gave him two weeks in which to pay the fine.

(a) The fine was given to him by the judge for two weeks.

(b) He was given the fine by the judge for two weeks.

(c) He was given two weeks by the judge in which to pay the fine.

(d) He was given two weeks by the judge for the fine.

67. Candidates who have failed will not be allowed to resit the examination.

(a) We cannot allow candidates to fail when they resit the examination.

(b) If candidates fail the examination they will have to resit it.

(c) The authorities will not allow failed candidates to resit examination.

(d) Resitting the examination is only possible for failed candidates.

68. One way of reducing tension is to learn to relax.

(a) Learning to relax can help reduce tension.

(b) If you want to learn to relax you should reduce tension.

(c) By reducing tension you will learn to relax.

(d) In order to relax you should reduce tension.

...5/-

69. John said it was amazing how similar I looked to his brother.

(a) "It's amazing how similar you look to your brother," said John.

(b) "It was amazing how similar you looked to my brother," said John.

(c) "It's amazing how similar you look to my brother," said John.

(d) "It was amazing how similar you looked to your brother," said John.

70. I dislike flying in the way that you dislike sailing.

(a) I don't like flying and neither do you.

(b) I don't like sailing or flying; nor do you.

(c) I like neither flying nor sailing, like you.

(d) I don't like flying and you don't like sailing.

71. Hydrogen, which has only one proton and one electron, is a light gas.

(a) Hydrogen which has only one proton and one electron is a light gas.

(b) Hydrogen has only one proton and one electron which is a light gas.

(c) Hydrogen is a light gas which has only one proton and one electron.

(d) Hydrogen has only one proton and one electron, which is a light gas.

72. I said, "Let's not jump to conclusions. Let's wait till we hear confirmation of this rumour."

(a) I told everyone not to jump to conclusions.

(b) I ordered everyone not to jump to conclusions.

(c) My advice was to wait till we heard confirmation of the rumour.

(d) My suggestion was to wait unless we heard confirmation of the rumour.

73. The disaster at Chernobyl is likely to rekindle doubts over the use of nuclear power.

(a) The disaster has made people think again about the use of nuclear power.

(b) The disaster will probably make people think again about the use of nuclear power.

(c) People have doubted the use of nuclear power after this disaster nuclear power.

(d) People will think again about the use of nuclear power after this disaster.

For the next 4 questions decide what kind of sentence or text you are reading.

74. A device used for keeping food fresh or frozen is called a refrigerator.

(a) classification (b) definition

(c) description (d) exemplification.

...6/-

75. Speech is a sound form of language which gives information by chopping up vocal sounds into segments.

(a) description (c) exemplification

(b) definition (d) classification

76. Yeast is added to paraffin in a tank along with water, air, ammonia and mineral salts. The yeast cells feed on the paraffin and start to grow. The yeast solution then goes into a centriguge where it is spin rapidly. The concentrated yeast, now thick and creamy, goes from the centrifuge to a container. It then passes into a drier where the cream is heated and the water evaporates. The purified yeast then appears as a fine powder. This yeast powder has a very high protein content. It does not have a very pleasant taste but could ultimately provide a very valuable food for man.

(a) description (c) exemplification

(b) definition (d) classification

77. Parasitology is the branch of biology which deals with the nature of parasitism and its effects on both the parasite and the host. Broadly speaking, a parasite is an organism which lives for all or part of its life on or in another organism from which it derives some benefit, such as food, shelter or protection. Organisms living on the host are known as extoparasites, these living within the host organism are called endoparasites.

(a) description (c) exemplification

(b) definition (d) classification

Look at the following sentences. When re-arranged in the correct order they make a logical paragraph.

A. The dry pulp sheets are turned into wet pulp.

B. The thin sheet is dried.

C. Paper is often made from dry pulp sheets.

D. The damp layers of pulp are pressed into a thin sheet.

78. Which is the first sentence?

(a) (b) (c) (d)

79. Which is the second sentence?

(a) (b) (c) (d)

80. Which is the third sentence?

(a) (b) (c) (d)

81. Which is the fourth sentence?

(a) (b) (c) (d)

Now look at the following sentences and do the same.

(a) Despite its wide distribution, carbon constitutes only 0.19 percent of the earth's crust.

(b) Carbon is a solid non-metallic chemical element occuring in the pure crystallire form as diamond ang graphite.

(c) It is also found in the combined form as a contituent of all organic materials, including coal and petroleum.

82. Which is the first sentence? (a) (b) (c)

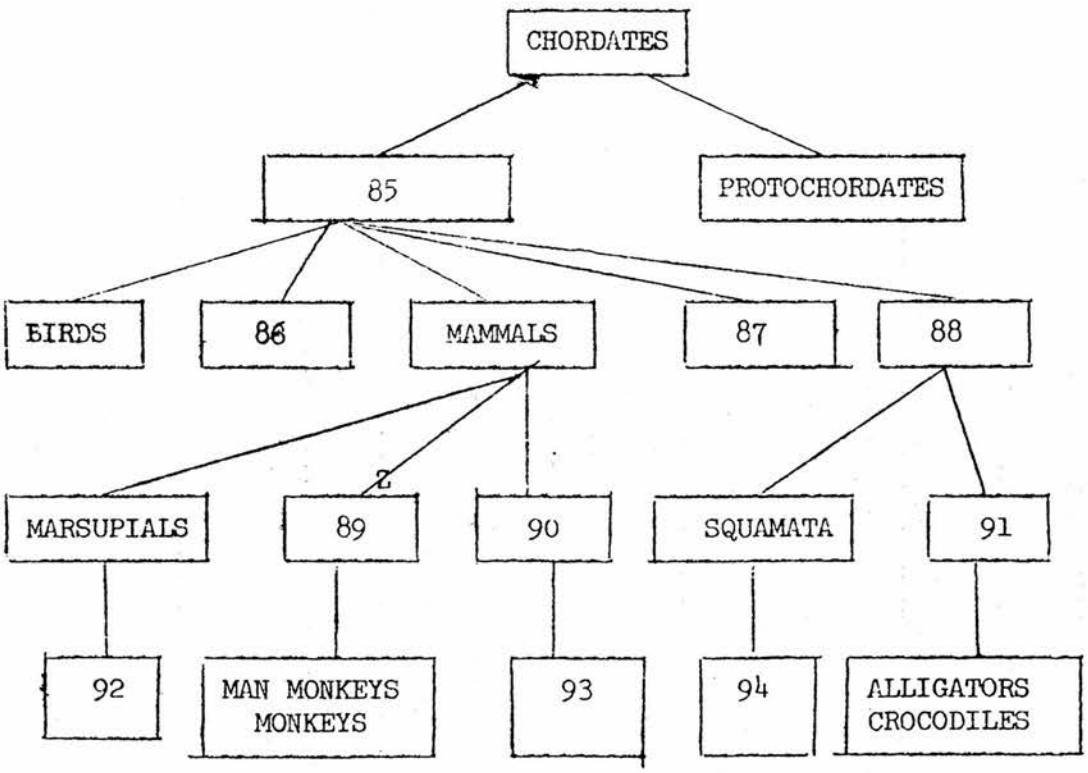
83. Which is the second sentence? (a) (b) (c)

84. Which is the third sentence? (a) (b) (c)

Look quickly at the following passage then complete the diagram with the appropriate word.

CHORDATES

The chordates are a large and highly diverse animal group which comprises vertebrates or animals with backbones (often referred to as the higher chordates) as well as a group of animals which lack vertebrae but which resemble vertebrates in other important respects. These are referred to as protochordates, or lower chordates. The vertebrates are divided into five classes: fishes, amphibians, reptiles, birds and mammals. Each of these five classes can be further subdivided into smaller groups: for example, mammals can be classified into 18 groups known as orders. Examples of orders are marsupials (such as kangaroos), primates (including man and the monkeys), and carnivores (including dogs and cats). The class of reptiles consists of five orders: examples of these are crocodilians (including crocodiles and alligators) and squamata, examples of which are snakes and lizards.



85. (a) chordates (c) backbones
(b) vertebrates (d) animals

86. (a) fishes (c) reptiles
(b) amphibians (d) mammals

87. (a) mammals (c) amphibians
(b) reptiles (d) fishes

88. (a) reptiles (c) fishes
(b) amphibians (d) mammals

89. (a) carnivores (c) crocodilians
(b) primates (d) reptiles

90. (a) mammals (c) primates
(b) crocodilians (d) carnivores

91. (a) carnivores (c) fishes
(b) primates (d) crocodilians

92. (a) dogs/cats (c) snakes/lizards
(b) kangaroos (d) carnivores

93. (a) dogs/cats (c) snakes/lizards
(b) kangaroos (d) carnivores

94. (a) dogs/cats (c) snakes/lizards
(b) kangaroos (d) carnivores

Now look at the following passage and fill in the blanks with the correct word from the list below.

Cultures have definite patterns. (95)___ these patterns are modified (96)_____ they are transmitted from one generation to the next.(97)_____ these changes take place slowly and sometimes they are rapid. (98)_____ the medieval era was for Western civilization a period of fairly slow change in culture patterns, (99)_____ the modern period has been characterized by rapid and dramatic changes. (100)_____-, in spite of these changes, a coherent pattern remains.

95. (a) Although (b) But (c) Since (d) Also
96. (a) as (b) during (c) also (d) so
97. (a) often (b) rarely (c) sometimes (d) usually
98. (a) Morever (b) In addition (c) Whereas (d) For example
99. (a) also (b) white (c) in addition (d) since
100. (a) However (b) Despite (c) Although (d) Indeed

UNIVERSITI SAINS MALAYSIA
PUSAT BAHASA DAN TERJEMAHAN

ENGLISH LANGUAGE PLACEMENT TEST
FORM C

PART 3 - 1½ hours

Part 3 is a test of your English comprehension. There are five passages, each with eight questions - a total of 40 questions.

Read through each passage in turn and then answer the questions on it.

Mark your answers - (a), (b), (c) or (d) - in thick pencil on the answer sheet.

Now turn over and begin Part 3. Work as quickly as possible.

...2/-

PASSAGE A

Child saved with help of famous writer of detective stories

By John Roper
Health Services Correspondent

A nurse who was reading an Agatha Christie thriller indirectly saved the life of a severely ill child whose condition baffled doctors at Hammersmith Hospital, London, it was learnt yesterday.

One Sunday morning a girl, aged 19 months, flown to England from Qatar, was admitted to the hospital semi-conscious and unresponsive to speech or commands.

All the resources of the hospital were used to establish a diagnosis, but doctors were at a loss.

The child's condition seemed to decline. Her blood pressure suddenly increased, she became more moribund, and the use of a respirator machine was considered. The decision was difficult because there was no firm diagnosis. Happily, her breathing improved spontaneously.

The next day, at the routine ward round, Marsha Maitland, the nurse with particular responsibility for the child, put down the book she was reading and interrupted the doctors' discussion with a suggestion that the child seemed to have thallium poisoning.

The doctors were surprised. Nurse Maitland said that in A Pale Horse, the Agatha Christie book she was reading, thallium poisoning was described, and the child's symptoms were remarkably similar. The one consistent feature emphasized in the book, loss of hair, seemed to be developing in the child that morning.

The doctors listened. Thallium poisoning was not one of the toxic substances screened in the laboratory tests, and the laboratory, in answer to a request, said that they were unable to carry it out.

Advice was sought from the London police who had the address of a laboratory which would test for thallium poisoning. The police told the doctors they had an expert living near them: Graham Young, serving life imprisonment in a jail next door to the hospital. The police said that Young had kept meticulous notes throughout his studies on the effects of thallium.

The laboratory test showed that the child's body contained more than ten times the permitted maximum of thallium. Inquiries from the child's parent suggested that the most likely source of the thallium, probably ingested by the child over a long period, was a domestic poison commonly used where she lived to kill cockroaches and rodents.

(From "The Times", London)

...3/-

101. Nurse Maitland was reading a book which described

(a) Agatha Christie

(b) a certain type of poisoning

(c) sick children

(d) hair care

102. People suffering from thalium poisoning begin to

(a) lose their hair

(b) feel surprised at their illness

(c) have low blood pressure

(d) forget things

103. After hearing Nurse Maitland's diagnosis, the doctors said they were unable to confirm it because

(a) they had already tested the child for thallium poisoning

(b) Thallium poisoning was not caused by toxic substances.

(c) they did not possess suitable equipment

(d) permission from the police was first needed to carry out experiments.

104. Graham Young was able to help the doctors because

(a) he had carefully studied cases of thallium poisoning

(b) he himself had poisoned the child

(c) he had the address of a laboratory which would test for thallium poisoning.

(d) he himself was also suffering from thallium poisoning.

105. The girl had probably consumed small quantities of

(a) cockroaches

(b) food infected by cockroaches & rodents

(c) insect and pest poison

(d) horse meat

.../-

106. The word "baffled" (line 1) means

(a) confused
(b) puzzled
(c) excited
(d) interested

107. "It" (line 25) refers to

(a) thallium poisoning
(b) the request
(c) the laboratory
(d) the laboratory test

108. "spontaneously" (line 13) means

(a) quickly
(b) by itself
(c) over a period of time
(d) easily

...5/-

PASSAGE B

ETHICAL DIMENSIONS OF THE ECOLOGICAL CRISIS

As with so many of the major problems of society, the precise extent and nature of the environmental crisis is not entirely clear. On the one hand are those who prophesy that humankind is facing global disaster in the near future. Our civilization as we know it will die or be disfigured beyond recognition unless we drastically change our ways. Grim predictions of potential global disaster are so widely broadcast that the present generation of young people have literally been weaned on these dire warnings. On the other hand are those who do have faith in the future. They submit that the human species is too great a biological success to end so abruptly and so soon. A species that can learn from the experiences of its predecessors, and in so doing fashion for itself a world unlike any experienced before, can continue to build new knowledge, achieving thereby still higher levels of attainment. Which view will be correct cannot be determined with certainty, at least not now.

Regardless of one's view, many people today are deeply distressed with the condition of both their social and natural environments. A large number of individuals have become apprehensive to the point of feeling threatened in a fundamental way. For example we do not know how many people this earth can provide for or at what level of existence, but we do know that there is a limit and we may be approaching it. We do not know how much wider the gap can grow between the rich and poor nations before Armageddon, but we do know there is a limit to how much suffering and oppression people can and will tolerate. We do not know the extent of the world's nonrenewable natural resources, but we do know the world is running out of gas. We do not know how much pollution this earth can absorb before it lashes back at its human antagonists, but we do know that the air is bad, sometimes the water may be unsafe to drink, and toxic substances are being let loose in the land (over 100,000 of them, according to a recent EPA report).

...6/-

109. How many different points of view are presented in the first paragraph?

(a) 1

(b) 2

(c) 3

(d) 4

110. The author suggests that young people today

(a) generally have a pessimistic view of the world

(b) have more faith in the future than their parents

(c) have become so used to warnings global disaster that they no longer listen to them.

(d) have grown up with constant warnings of global disaster

111. The argument of these who have faith in the future is based on the fact that the human species

(a) is and always has been biologically adaptable

(b) has learned from the experiences of its predecessors

(c) has the ability to fashion the world for itself

(d) can use new knowledge to achieve higher levels of attainment.

112. Many people feel apprehensive and threatened because

(a) they don't have a secure world view

(b) they are very worried about the condition of their social and natural environments

(c) they are distressed at the number of people crowded into the Earth.

(d) they feel that the end of the world is near.

113. The sentence beginning "We do not know how much pollution ..." (line 23) is

(a) an argument for the main idea of the passage

(b) an example of the main idea of paragraph 2

(c) the main idea of paragraph 2

(d) the topic sentence of the whole passage

114. The phrase "Regardless" of one's view" (line 14) here means

(a) Whether one has a view or not
(b) If one has a view
(c) Even though one doesn't have a view
(d) Whatever one's view

115. "antagonists" (line 25) means

(a) enemies
(b) opposites
(c) populations
(d) problems

116. "it" (line 19) refers to

(a) this earth
(b) level of existence
(c) a limit
(d) the gap

...8/-

PASSAGE C

Smoking and cancer

Americans smoke six thousand million cigarettes every year (1970 figures). This is roughly the equivalent of 4,195 cigarettes a year for every person in the country of 18 years of age or more. It is estimated that 51% of American men smoke compared with 34% of American women.

Since 1939, numerous scientific studies have been conducted to determine whether smoking is a health hazard. The trend of the evidence has been consistent and indicates that there is a serious health risk. Research teams have conducted studies that show beyond all reasonable doubt that tobacco smoking, particularly digarette smoking is associated with a shortened life expectancy.

Cigarette smoking is believed by most research workers in this field to be an important factor in the development of cancer of the lungs and cancer of the throat and is believed to be related to cancer of the bladder and the oral cavity. Male cigarette smokers have a higher death rate from heart disease than non-smoking males. (Female smokers are thought to be less affected because they do not breathe in the smoke so deeply.) The majority of physicians and researchers consider these relationships proved to their satisfaction and say, 'Give up smoking. If you don't smoke - don't start!'

Some competent physicians and research workers - though their small number is dwindling even further - are less sure of the effect of cigarette smoking on health. They consider the increase in respiratory diseases and various forms of cancer may possibly be explained by other factors in the complex human environment - atmospheric pollution, increased nervous stress, chemical substances in processed food, or chemical pesticides that are now being used by farmers in vast quantities to destroy insects and small animals. Smokers who develop cancer or lung diseases, they say, may also, by coincidence, live in industrial areas, or eat more canned food. Gradually, however, research is isolating all other possible factors and proving them to be statistically irrelevant.

Apart from statistics, it might be helpful to look at what smoking tobacco actually does to the human body. Smoke is a mixture of gases, vaporized chemicals, minute particles of ash, and other solids. There is also nicotine, which is a powerful poison, and black tar. As the smoke is breathed in, all these components form deposits on the membranes of the lungs. One point of concentration is where the air tube, or bronchus, divides. Most lung cancer begins at this point.

Smoking also affects the heart and blood vessels. It is known to be related to Beurger's disease, a narrowing of the small veins in the hands and feet that can cause great pain and lead even to amputation of limbs. Smokers also die much more often from heart disease.

While all tobacco smoking affects life expectancy and health, cigarette smoking appears to have a much greater effect than cigar or pipe smoking. However, nicotine consumption is not diminished by the latter forms, and current research indicates a causal relationship between all forms of smoking and cancer of the mouth and throat. Filters and low tar tobacco are claimed to make smoking to some extent safer, but they can only marginally reduce, not eliminate the hazards.

...9/-

117. According to 1970 figures, in the US

(a) these are twice as many men smokers as women smokers

(b) over half the adult population are smokers

(c) About 43% of the total population are smokers

(d) Just under half the adult population are smokers.

118. The evidence since 1939 shows that

(a) there is consistent but not absolute evidence that smoking is a serious health risk

(b) male cigarette smokers have a shorter life expectancy than female smokers

(c) nonlsmoking males and smoking females have about the same risk of developing lung cancer

(d) smoking is the most important factor in the development of lung cancer

119. Most physicians and research workers believe that respiratory diseases and cancers may have increased because

(a) of atmospheric pollution

(b) of cigarette smoking

(c) the human environment is complex

(d) the people affected live in industrial areas and eat more canned food

120. The logical implication of the three sentences beginning "As the smoke ...." (lines 34 - 37) is that

(a) lung cancer is definitely caused by smoking

(b) lung cancer is unlikely to be caused by smoking

(c) lung cancer is probably caused by smoking

(d) lung cancer and smoking just happen to affect the same part of the body

121. Life expectancy and health

(a) are reduced more by smoking cigars or pipes than by cigarettes

(b) can be returned to normal by smoking filter cigarettes

(c) are affected more by cigarettes than by cigars or pipes

(d) depend on the quantity of nicotine consumed

PASSAGE D

Stop that noise!

Noise can be defined simply as unwanted sound, It is immediately obvious that it is very subjective in that the latest full quadrophonic sound reproduced on a newly acquired set can be most delightful to the connoisseur of music but plain murder to the tired night-shift worker next door trying to sleep.

Unwanted sound or noise is usually received at one of three levels: (a) the tolerance level, (b) the annoyance level, and (c) the painful level or the level where actual damage occurs.

People have different levels of noise tolerance and this is also affected by the environment in which they find themselves. Noise in the first category could range from the background whine of fans and air-conditioners to that experienced by a typist in a typing-pool. The noise in a typing pool could well be put in the second classification by a visitor who is not used or "conditioned" to the continuous sound of type falling on paper. Nevertheless, even to the typist herself, working in a noisy environment contributes to the stress of her occupation and she will probably feel a sense of relief when she returns to quieter surroundings.

Annoyance is also often caused by an intermittent impulsive sound such as doors slamming. Here again the environment is important as the sound of doors slamming in a noisy office is less annoying than if it was a quiet residence.

Noise levels in the last category, the level where actual damage results, are not likely to occur in the home or office. Hearing loss due to a "burst" eardrum is rare and is usually a gradual process where the delicate hair cells in the cochlea of the ear are occasionally damaged by loud noises and not replaced. Nevertheless, those who frequent noisy discotheques often experience at least a temporary deafness on emergence. It is also a fact that personnel working in very noisy environments without adequate hearing protection have gone deaf in middle life.

It is common knowledge that hearing deteriorates with age but it is now known that this condition is accelerated by long periods of exposure to loud sounds.

Fortunately, it is possible to reduce noise in our homes by keeping in mind some very simple rules when we furnish our homes. If a room is air-conditioned, for example, a worthwhile improvement can be obtained for a moderate additional outlay by the use of a special double-panelled glass for doors and windows. For most people, however, it is still possible to keep windows open and reduce the noise entering the room by large, thick curtains. False curtains across the walls of a room can also help reduce noise. Noise-absorbent tiles on walls and ceilings, and carpets on floors are a familiar sight in most offices and will likewise help noise reduction in the home.

There are several other ways in which noise can be reduced in the home. In the final analysis, however, each person must behave as a responsible member of the community and try to reduce the noise he causes.

...12 -

130. Large thick curtains are useful because

(a) they enable the inhabitants of a flat to lead a private life

(b) they calm people who are annoyed by constant noise

(c) they help to keep out unwanted noise

(d) they improve the decoration of a room

131. "it" (line 22) refers to

(a) an intermittent impulsive sound

(b) the environment

(c) a noisy office

(d) the sound of doors slamming

132. "it" (line 2$) refers to

(a) personnel ... have gone deaf in middle life.

(b) adequate hearing protection

(c) a temporary deafness

(d) deafness in middle life.

...14/-

133. According to the passage, the justification of a university is:

(a) it imparts information to young and old

(b) it present facts and experience to young and old

(c) it imparts knowledge to imaginative people

(d) it combines imagination with knowledge and experiance

134. In the author's opinion, imagination

(a) applies to general principles as well as to facts

(b) is an intellectual survey of alternative possibilities

(c) allows us to build an intellectual vision of a new world.

(d) should be clearly separated from facts.

135. According to the author, youth

(a) has the energy of imagination

(b) needs to be strengthened by discipline

(c) has a great measure of strength

(d) is imaginative but needs knowledge and experience.

136. The sentence beginning "At least ..." (line 4)

(a) provides the result of the topic sentence

(b) provides the reason for the topic sentance

(c) adds a minimum requirement for the topic sentence

(d) adds support to the topic sentence

137. The last two sentences in paragraph 1 (lines 6 - 8)

(a) clarify the topic sentence by adding examples

(b) provide a new idea which will be developed later

(c) add additional functions that the university should fulfill.

(d) provide a fuller definition of "imagination"

122. "dwindling" (line 21) means

(a) insignificant
(b) increasing
(c) decreasing
(d) investigating

123. "processed" (line 25) means

(a) factory treated
(b) hygienically packed
(c) stared
(d) constructed

124. "it" (line 31) refers to

(a) statistics
(b) research
(c) to look at what smoking ..... does
(d) smoking tobacco

125. If noise is described as being at the tolerance level, it

(a) can be harmful to the health of people

(b) is annoying to some people

(c) is accepted by some people without complaint

(d) takes the form of a continuous sound but is not too loud

126. If two people live or work in different environments a certain noise

(a) will have the same effect on both

(b) may be tolerated by one person but cause the other person to be annoyed.

(c) will probably be painful to one of them

(d) may nevertheless cause stress to both to an equal degree.

127. A person who works in noisy surroundings may tolerate noise

(a) but may feel stress as a result of it

(b) and may not wish to return to quieter surroundings

(c) but may be relieved to hear that others find the same surroundings noisy.

(d) and may even think that it contributes to his or her work

128. From the text it can be assumed that the banging of doors is not as annoying in a noisy office as it is in

(a) a library

(b) an airport

(c) a discotheque

(d) a market

129. Serious damage to the ears

(a) is not usually caused suddenly

(b) only occurs where the hair cells are delicate

(c) frequently results from noise at the annoyance level

(d) sometimes takes the form of temporary deafness.

..13 /-

138. The word "discipline" (line 15) means

(a) instruction and exercise designed to train
(b) a branch of learning
(c) punishment designed to change behaviour
(d) a subject studied at university

139. The word "this" (line 4) refers to

(a) imparting information imaginatively
(b) the university
(c) the reason for the university's existance
(d) the imaginative consideration of learning.

140. "it" (line 13) refers to

(a) a new world
(b) an intellectual vision
(c) imagination
(d) an intellectual survey

PASSAGE E

Universities and Imagination

The justification for a university is that it preserves the connection between knowledge and the zest of life, by uniting the young and the old in the imaginative consideration of learning. The university imparts information, but it imparts it imaginatively. At least, this is the function which it should perform for society. A university which fails in this respect has no reason for existence. This atmosphere of excitement, arising from imaginative consideration, transforms knowledge. A fact is no longer a burden on the memory: it is energizing as the poet of our dreams, and as the architect of our purposes.

Imagination is not to be divorced from the facts: it is a way of illuminating the facts. It works by eliciting the general principles which apply to the facts, as they exist, and then by an intellectual survey of alternative possibilities which are consistent with those principles. It enables men to construct an intellectual vision of a new world, and it preserves the zest of life by the suggestion of satisfying purposes.

Youth is imaginative, and if the imagination be strengthened by discipline, this energy of imagination can in great measure be preserved through life. The tragedy of the world is that those who are imaginative have but slight experience, and those who are experienced have feeble imaginations. Fools act on imagination without knowledge; pedants act on knowledge without imagination. The task of a university is to weld together imagination and experience.

ALRED NORTH WHITEHEAD

UNIVERSITI SAINS MALAYSIA
PUSAT BAHASA DAN TERJEMAHAN

ENGLISH LANGUAGE PLACEMENT TEST
FORM C

PART 4 - ½ hour

Part 4 is a test of your ability to summarize a passage of English

Read quickly through the passage on the next page, then write a summary of the passage in English.

You should use between 80 - 150 words.

Use the separate answer sheet for your answer and for your rough work.

...2/-

DIGITAL COMPUTERS AND THEIR USES

In the digital computer the numbers to be manipulated are represented by sequences of digits which are first recorded in suitable code - usually the binary code -, are then converted into positive and negative electrical impulses, and stored in electrical or magnetic registers which serve basically the same purpose as the counting wheels in a desk calculating machine.

The technique of making the computer carry out a particular calculation is known as 'programming', which involves first breaking the calculation down into a sequence of arithmetic operations, and then preparing a series of instructions which cause the computer to carry out the required operations on the stored information in the correct order. It is now possible to add or subtract two large numbers in one to two microseconds, and to multiply or divide them in ten to twenty microseconds, so that a computer can perform as much arithmetic in a quarter of an hour as an efficient clerk with pencil and paper might reasonably hope to achieve in a lifetime.

There are many situations in which this ability to handle and to analyse large quantities of arithmetic data according to instructions is of great value. Some examples are fileds of scientific investigation such as crystallography, atomic physics and astronomy, where masses of experimental data are involved and complex theoretical concepts need to be tested against them; in engineering design where the design parameters , of which there are many, can be varied systematically and their effects studied and optimized; and for the storage of reference data in libraries and insurance offices in such a way as to afford ready access to particular references on request.

A particularly important application of the digital computer in simplified form is as a component in the control equipment of manufacturing processes - as the nerve centre which accumulates and analyses data recording the operating conditions and performances of the plant, and sends out instructions, when appropriate, for their modification. This is one aspect of what is called 'automation' - the replacement of human control by instrumental control. The completely automatic factory is no longer a fantasy. What is restraining its realization are the difficulties in handling the severe economic considereations and the cmplex human problems involved.

During a recent conference on 'Computable Models in Decision-taking , the chairman said:

> 'It is significant that despite (the rapid advance of science since the seventeenth century, it made no impact on the problems of prediction until the advent of the digital computer, and it is only thanks to (... the powerful new computers that worth-while prediction in human affairs has been possible at all.'

He went on to say that accurate prediction, and therefore decision-taking, is only possible in autonomous systems for which the laws of behaviour are known. This is not, of course, the case in human affairs, and the most that can be yet done is to seek for mathematical models which describe, however imperfectly as yet , the presumed behaviour of the system or situation under investigation, and then to study systematically,

.../2/-

with the aid of the computer, the consequences which arise from the variation of the parameters invorporated in the model. It was fascinating to hear of some of the problems falling within the shpere of human behaviour to which the computer is now being applied experimentally. One was that of forecasting the sale of particular kinds of fabric, with a view to ensuring that out-of-stock situations do not develop; another, predicting the future demand for various categories of steel products, as a basis for judging the necessary provision of manufacturing capacity; and a third, an attempt to establish a model of the economic system of the country, as a means of predicting its future pattern of development.

The point that I am anxious to make is that the search for models of this kind, the study of their behaviour and of the relationship of this behaviour to the real situations which they seek to represent, and the consequential modification of them so as to lead to reliable prediction and then to decision-taking, would not be possible were it not for the assistance afforded to the investigator by the digital computer - and by the work of the technologists who have successfuly transformed the scientific ideas on which it is founded into stable, reliable and economic pieces of electrical equipment.

Where this new tool of investigation will ultimately take us is beyond my powers of prediction. But the subject is only a few years old and with the improvements in electronic techniques which may confidently be expected, and the rapidly increasing knowledge and understanding of the brain possessed by the medical profession, it would perhaps be unwise to forecast undue restrictions on the nature of the ultimate achievement.

JACKSON, SIR WILLIS, Penguin Technology Survey 1966,
Penguin Books Ltd.

UNIVERSITI SAINS MALAYSIA
PUSAT BAHASA DAN TERJEMAHAN

EENGLISH LANGUAGE PLACEMENT TEST
FORM D

PART 1 - (25 minutes)

Part 1 is a test of your knowledge of basic English grammar. There are 50 questions.

For each question choose the answer - (a), (b), (c), or (d) - which best completes the sentence.

Mark your answer in thick pencil on the answer sheet.

For example:

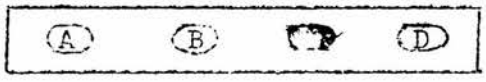
Question 104

This camera _____ good pictures.

(a) sees
(b) makes
(c) takes
(d) looks at

The correct answer is (c) - This camera takes good pictures.

You mark your answer sheet like this.

104 

Now turn over and begin Part 1. Work as quickly as possible.

...2/-

1. It is better _____ your money in a bank than under the bed.
   (a) to put (b) putting
   (c) by putting (d) put

2. _____ he was angry, he listened patiently.
   (a) Nevertheless (c) In spite of
   (b) But (d) Though

3. If I had a lot of money I suppose _____ a house.
   (a) I'd buy (c) I'll buy
   (b) I have bought (d) I'm buying

4. Someone _____ my watch while I was asleep.
   (a) need have stolen (b) should have stolen
   (c) must have stolen (d) didn't need to steal

5. I can't read Greek, so _____ the documents translated.
   (a) I'm having (c) I'll do
   (b) I'm doing (d) I'll make

6. The car crashed into a queue of people _____ 4 were killed.
   (a) where (c) of whom
   (b) so (d) by which

7. That boy _____ his face, he likes being dirty.
   (a) can't wash (c) is going to wash
   (b) won't wash (d) will wash

8. After much effort they managed _____ the top of the mountain.
   (a) in reaching (b) by reaching
   (c) to reach (d) and reached

9. He lives somewhere _____ France at the moment.
   (a) after (c) at
   (b) over (d) in

10. If I had _____ idea I would help you immediately.
    (a) the (b) any
    (c) _____ (d) no

11. Long hours of study don't worry me because I _____ hard
    (a) used to working (b) am used to working
    (c) am used to work (d) used to work

...3/-

12. If you_____ my advice you wouldn't be in this mess now.
(a) had taken (c) take
(b) took (d) have taken

13. A surgeon is _____ a dentist in one respect: both need manual dexterity .
(a) same (b) like
(c) as (d) different from

14. This is a difficult book; _____ it is worth studying carefully.
(a) on the other hand (b) on the contrary
(c) in addition (d) and so

15. After _____ doctor had examined the wound, he put a bandage on it.
(a) the (c) some
(b) a (d) _____

16. A land breeze is produced at night _____ a sea breeze is produced during the day.
(a) yet (c) moreover
(b) while (d) as well as

17. As soon as she _____ to type I'll give her a job.
(a) learns (b) would learn
(c) will learn (d) is learning

18. He doesn't eat meat because he's _____ vegetarian
(a) a (b) the
(c) some (d) any

19. We shall not succeed _____ we get the staff and the equipment.
(a) if (c) while
(b) until (d) by the time that

20. He lives _____ a small village near Cambridge
(a) on (b) at
(c) in (d) by

21. If the EXIT doors hadn't been blocked, people _____ to escape.
(a) were able (b) would be able
(c) would have been able (d) are able

22. This pocket calculator is _____ then that one
(a) reliable (b) most reliable
(c) more reliable (d) the more reliable

23. Experts in kinesics are not prepared _____ a precise vocabulary of gestures.

(a) to spell out (c) for spelling out
(b) spelling out (d) at spelling out

24. A bookshop sells books, _____ a library only lends them.

(a) on the contrary (b) moreover
(c) in addition (d) whereas

25. John, _____ father works abroad, has nearly finished his studies.

(a) whose (b) who
(c) whom (d) of whom

26. The colour has changed, so the substance _____

(a) could have dissolved (c) must have dissolved
(b) is dissolving (d) needs to have dissolved

27. Romeo and Juliet were two lovers _____ parents hated each other.

(a) whose (c) their
(b) which (d) and

28. A concerto is _____ piece of music for orchestra and solo. instrument.

(a) the (b) some
(c) a (d) _____

29. Penicillin has harmful side effects; _____ sulpha.

(a) so (c) likewise
(b) in addition (d) indeed

30. I hope you are enjoying your visit; _____ any new friends yet?

(a) did you make (b) are you making
(c) have you made (d) do you make

31. Back muscles can be made more flexible _____ Hatha Yoga.

(a) if practising (c) by practising
(b) with practising (d) to practice

32. I'm sorry about the mess but I _____ my house re-wired.
(a) have (b) do
(c) am making (d) am having

33. Of all the goods in the shop that one is_____ .
(a) more expensive (b) the most expensive
(c) expensive (d) the expensive one

34. A person _____ specialises in the study of weather patterns is called a meteorologist.
(a) who (b) whose
(c) which (d) of whom

35. Provided you _____ the instructions you should be able to operate the machine.
(a) will follow (b) would follow
(c) had followed (d) follow

36. I know the strike is annoying but we have to put _____ .
(a) in for (b) up with
(c) out of (d) up to

37. It _____ heavily when he woke up.
(a) rained (b) was raining
(c) is raining (d) has rained

38. A bridge is a structure _____ is to span a gap.
(a) which (c) which purpose
(b) whose purpose (d) what

39. An adjective modifies a noun _____ an adverb modifies a verb.
(a) whereas (c) on the contrary
(b) moreover (d) yet

40. When I hear from him _____ you know.
(a) I'll let (c) I let
(b) I'm going to let (d) I have let

...6/-

41. After he _____ I began to worry about him.
(a) was going (c) had gone
(b) has gone (d) goes

42. It's a serious injury, but he _____ again in six weeks.
(a) walks (b) will be walking
(c) would walk (d) is walking

43. What _____ if I press this button?
(a) had happened (c) would happen
(b) will happen (d) happened

44. The car is _____ useful invention.
(a) the (c) an
(a) a (d) ____

45. They succeeded _____ escaping from the fire.
(a) to (b) by
(c) with (d) in

46. They were able to wade across, so they _____
(a) needn't swim (c) must have swim
(b) didn't need to swim (d) must not swim

47. I _____ him for three months now.
(a) haven't seen (c) am seeing
(b) didn't see (d) had seen

48. The teacher usually _____ lectures once a week
(a) gives (b) is giving
(c) has given (d) was giving

49. If you don't know the meaning, look it _____ in a dictionary.
(a) over (c) up
(b) after (d) on

50. He said _____ here but I can't see him.
(a) he will be (c) would
(b) he has been (d) he would be

UNIVERSITI SAINS MALAYSIA
PUSAT BAHASA DAN TERJEMAHAN

ENGLISH LANGUAGE PLACEMENT TEST
FORM D

PART 2 - (35 minutes)

Part 2 is a test of your working and reading grammar of English. There are 50 questions.

Not all the questions are of the same type, but you should in every case mark your answer - (a), (b), (c), or (d) - in thick pencil on the answer sheet.

For example:

Which sentence is closest in meaning to the first sentence?

105. John likes neither tea nor coffee

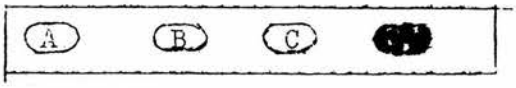
(a) John likes both tea and coffee

(b) John likes tea but not coffee

(c) John likes coffee but not tea

(d) John doesn't like coffee or tea

The correct answer is (d) - John doesn't like coffee or tea.
You mark your answer sheet like this:

105. 

Now turn over and begin Part 2. Work as quickly as possible.

51. Dr. Jones was advised by Dr. Smith to stay in bed for a few days.

(a) Dr. Smith had to stay in bed for a few days, on Dr. Jones' advice.

(b) It was Dr. Jones who stayed in bed after advising Dr. Smith.

(c) Dr. Jones advised him, and Dr. Smith agreed, to stay in bed for a few days.

(d) Dr. Smith's advice was that Dr. Jones should stay in bed for a few days.

52. Since the boy handled the rope skilfully, the kite rose higher and higher.

(a) Every time the boy handled the rope skilfully, the kite rose higher.

(b) The kite rose higher and higher due to the boy's skilful handling of the rope.

(c) Because the kite rose higher and higher, the boy handled the rope skilfully.

(d) In order to make the kite rise higher and higher, the boy handled the rope skilfully.

53. It looks as if food will soon be cultivated under the sea.

(a) We think that soon food may possibly be cultivated under the sea.

(b) We think that food is unlikely to be cultivated under the sea.

(c) The cultivation of food under the sea is an unlikely prospect.

(d) The cultivation of food under the sea is now a realistic possibility.

54. Whereas Singapore has many tall buildings, Johore Bharu has only a few.

(a) Unlike Singapore, Johore Bahru has several tall buildings.

(b) There are many tall buildings in Singapore, but not many in Johore Bahru.

(c) In contrast to Johore Bahru, Singapore has a few tall buildings.

(d) By comparison with Singapore, Johore Bahru has a few tall buildings.

55. The library, which is in High Street, is very well stocked.

(a) There is only one library and it is well stocked.

(b) There is only one library in High Street, there are others elsewhere.

(c) There is more than one library, but the one is High Street is well stocked.

(d) There is more than one library; one is in High Street, the others are elsewhere.

...3/

56. Prior to his return he had meant to throw it away.

(a) His intention was to throw it away before he returned.

(b) He intended to return and then throw it away.

(c) On his return he intended to throw it away.

(d) When he had returned, his intention was to throw it away.

57. "Leave the house and never darken my doorstep again," said the angry father to his son.

(a) The angry father told his son to leave the house and never darken his doorstep again.

(b) The angry father asked his son to leave the house and never darken his doorstep again.

(c) The angry father ordered his son to leave the house and never darken my doorstep again.

(d) The angry father instructed his son to leave the house and never to darken my doorstep again.

58. Shelters have been built in case of war breaking out.

(a) Due to the outbreak of war, shelters have been built.

(b) Shelters have been built in order to prevent war breaking out.

(c) The outbreak of war has led to the building of shelters.

(d) If war breaks out, shelters will have already been built.

59. It isn't necessary to buy a licence for a bicycle.

(a) You needn't buy a licence for a bicycle.

(b) You shouldn't have a licence for a bicycle.

(c) You mustn't have a licence for a bicycle.

(d) You don't have a licence for a bicycle.

60. Tropical rain forests are being chopped down rapidly.

(a) People have chopped down the rain forests.

(b) People are chopping down the rain forests.

(c) Rapid chopping will clear the rain forests.

(d) Clearing the rain forests has been done by chopping.

61. You can test acids and alkalis by means of Litmus paper.

(a) You can use acids and alkalis to test Litmus paper.

(b) Acids and alkalis are tested by means of Litmus paper.

(c) You can use litmus paper to test acids and alkalis.

(d) Litmus paper can distinguish between acids and alkalis.

...4/-

62. Here is my name and address. You may want to get in touch with me.

(a) Here is my name and address unless you want to get in touch with me.

(b) Here is my name and address should you want to get in touch with me.

(c) Here is my name and address lest you want to get in touch with me.

(d) Here is my name and address so that you can get in touch with me.

63. Although it rained heavily in town it was only drizzling here.

(a) It rained heavily both here and in town.

(b) It rained heavily in town; moreover it was only drizzling here.

(c) It didn't rain heavily in town; neither did it here.

(d) It rained heavily in town, whereas it was only drizzling here.

64. Brazil may well be independent of oil imports by 1990

(a) We hope that Brazil will be independent of oil imports by 1990.

(b) Brazil will be independent of oil imports by 1990.

(c) 1990 should see Brazil's independence of oil imports.

(d) Brazil can safely be predicted to be independent of oil imports by 1990.

65. Once you have rolled the pieces of dough you should shape them.

(a) Shape the pieces of dough then roll them.

(b) Roll the pieces of dough then shape them.

(c) Having shaped the dough, you should roll it.

(d) Roll the dough, after having shaped it.

66. As fossil fuels may soon be exhausted we should look for new supplies of energy.

(a) Fossil fuels may soon be exhausted; however, we should look for new supplies of energy.

(b) Fossil fuels may soon be exhausted; hence we should look for new supplies of energy.

(c) Fossil fuels may soon be exhausted; meanwhile we should look for new supplies of energy.

(d) Fossil fuels may soon be exhausted; moreover we should look for new supplies of energy.

67- My uncle who lives in England is very rich.

(a) I have more than one rich uncle and one of them lives in England.

(b) I have one uncle, and he lives in England and is rich.

(c) I have one rich uncle, and he lives in England.

(d) I have more than one uncle, one of them lives in England and is rich.

...3/-

68. He ran out of money and had to look for a job.

(a) He had to look for a job because he had run out of money.

(b) He had to look for a job, then he ran out of money.

(c) While looking for a job he ran out of money.

(d) Because he had to look for a job, he ran out of money.

69. He said that if it got cold I was to give him another blanket.

(a) "If you get cold, give me another blanket," he said.

(b) "If it gets cold give him another blanket," he said.

(c) "If it got cold I'd give you another blanket," he said.

(d) "If you get cold, I'll give you another blanket," he said.

70. "Get out of my way," he said.

(a) He said to me to get out of his way.

(b) He said to me to get out of my way.

(c) He told me to get out of his way.

(d) He told me to get out of my way.

71. This speed limit is to be introduced gradually.

(a) They will introduce this speed limit gradually.

(b) This speed limit which will be introduced-is very low.

(c) The gradual introduction of the speed limit will be done soon.

(d) The gradual introduction will lower the speed limit.

72. He coughed to warn them that he was coming.

(a) After coughing, he warned them, that he was coming.

(b) In order to warn them that he was coming, he coughed.

(c) As a result of coughing, he warned them that he was coming.

(d) Having warned them that he was coming, he coughed.

73. Unless you put on the brakes the car won't stop.

(a) Provided you put on the brakes the car won't stop.

(b) If you put on the brakes the car won't stop.

(c) The car will stop unless you don't put on the brakes.

(d) The car will stop provided you put on the brakes.

...6/-

For the next 4 questions decide what kind of sentence or text you are reading.

74. In general, geology is divided into the fields of physical and historical geology.

(a) description
(b) definition
(c) exemplification
(d) classification

75. Sounds fall into several groups: Those which can be tolerated, those which annoy, and those which cause actual damage.

(a) classification
(b) definition
(c) description
(d) exemplification

76. People are continually engaged in some learning activity or other - learning to ride a bicycle or speak a foreign language, to dance, swim, play a card game, handle a pneumatic drill, manage a shop or administer a government department. How is it that we can use the word 'learning' about such a varied set of activities? The only similarity lies in the fact that in each case there is a change in the learner brought about in some way by the interaction of the environment with the individual. If we adopt as a provisional definition of an instance of learning any more or less permanent change of behaviour which is the result of experience we find that even the most primitive animals are capable of some learning. In fact, in a very special sense, it can also be said that plants are able to learn.

(a) classification
(b) definition
(c) description
(d) exemplification

77. Micro organisms were once regarded as being members of the plant Kingdom, apart from protozoa which were classed as animals. It became obvious that this arbitrary classification resulted in confusions, even absurdities. A virus infecting an animal cannot by any criteria, be termed a plant. There became almost as many systems of classification as there were microbiologists. In order to clarify the nature of micro-organisms, we may distinguish between those, like fungi and some algae, which have a cell structure similar to higher organisms, and those, like the bacteria and the blue-green algae, which have a comparatively simple cell structure. We will refer to the former as "higher protists" and to the latter as "lower protists". Both these groups are placed in the Kingdom Protista. The viruses and the recently described subviral agents cannot at present be adequately classified, so we shall place them in a group of their own.

(a) classification
(b) definition
(c) description
(d) exemplification

Look at the following sentences. When re-arranged in the correct order they make a logical paragraph.

(a) There are two main types of programmes - linear and branching

(b) The value of teaching machines depends very much on the value of the programmes they contain.

(c) When the learner has mastered one step, he goes to the next.

(d) A linear programme consists of a series of small steps.

78. Which is the first sentence? (a) (b) (c) (d)

79. Which is the second sentence? (a) (b) (c) (d)

80. Which is the third sentence? (a) (b) (c) (d)

81. Which is the fourth sentence? (a) (b) (c) (d)

Now look at these sentences and do the same.

(a) This allows the chemicals to mix and produce steam.

(b) A second series of values is opened.

(c) As a result, the turbines start burning.

82. Which is the first sentence? (a) (b) (c) (d)

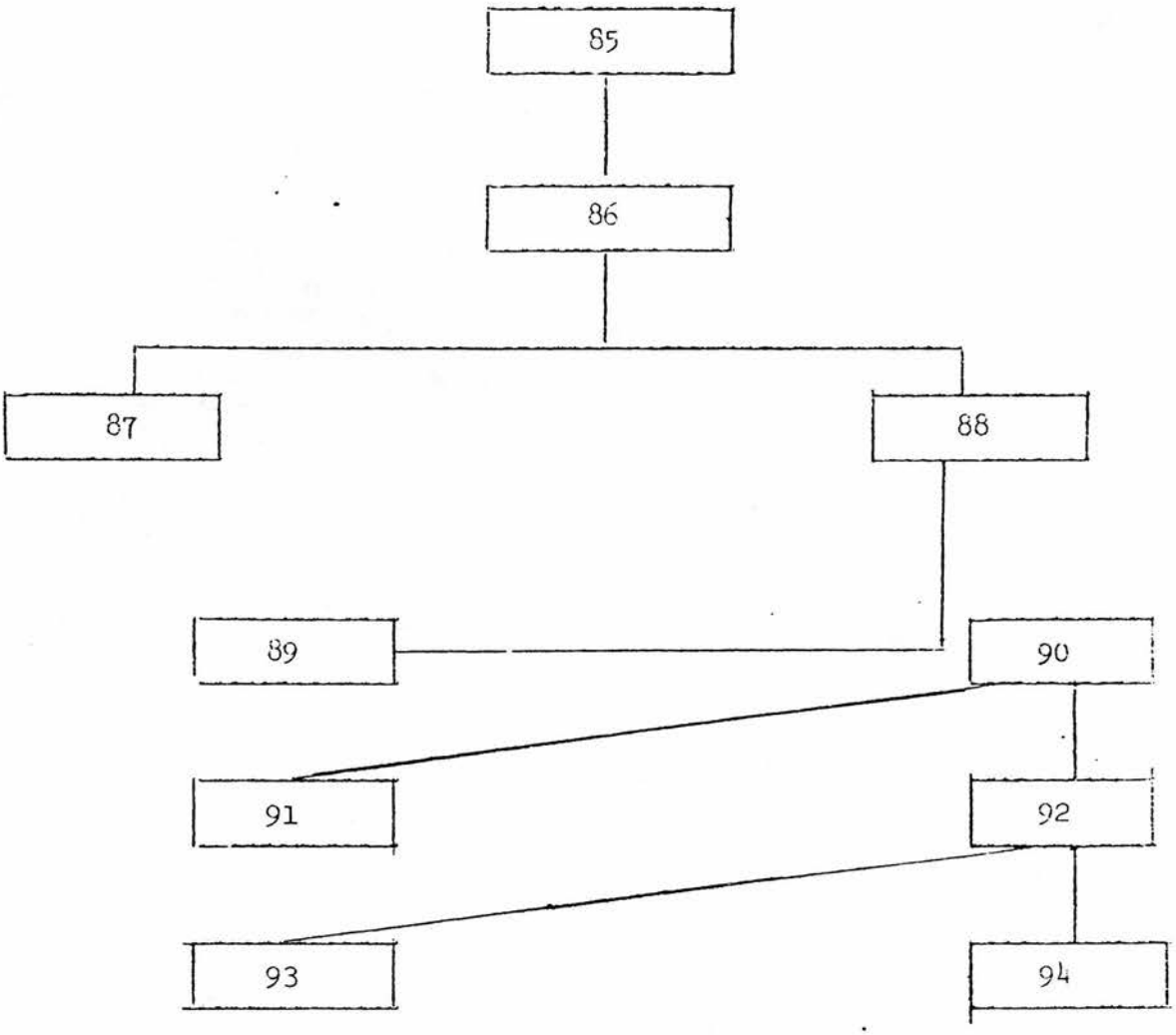
83. Which is the second sentence? (a) (b) (c) (d)

84. Which is the third sentence? (a) (b) (c) (d)

Read quickly through the following passage and then complete the diagram.

Behaviourist or Stimulus-Response theories have been extremely influential. The simplest type of S-R learning is usually described under the heading 'conditioning'. It is convenient to separate conditioning into two broad classes: classical and instrumental conditioning. The essential operation in classical conditioning is the pairing of two stimuli, as a result a of which the first stimulus elicits the response previously elicited by the second. The Pavlovian experiment in which dogs learned to salivate to the sound of a bell which had been paired repeatedly with food presentations has been taken to be the prototype of classical conditioning. Instrumental conditioning (or operant conditioning) is an experimental procedure in which an animal is given reinforcement after it spontaneously makes a particular response, and the intensity of the response then increases. Instrumental conditioning experiments can be classified according to the type of reinforcement into those using rewards and those using punishments. Both reward training and omission training use rewards. The former type of experiment is illustrated by the famous 'Skinner box'. There are two main kinds of experiments using punishment: escape training and avoidance conditioning. The latter can be subdivided into passive and active avoidance. In avoidance conditioning the animal is given, for example, a mild electric shock. Then a warning signal such as a red light is given, which enables the animal to avoid the shock. In active avoidance the animal makes a specific response, such as going into a different compartment in order to avoid the shock. In passive avoidance the animal must learn not to go into the compartment where there are shocks.

...8/-



85. (a) S-R learning (b) behaviourist theories
(b) learning (d) conditioning

86. (a) conditioning (b) S-R learning
(c) behaviourist theories (d) learning

87. (a) instrumental (b) conditional
(c) classical (d) behaviourist

88. (a) classical (b) behaviourist
(c) conditional (d) instrumental

89. (a) use of rewards (b) use of punishment
(c) pairing (d) stimuli

90. (a) use of rewards (b) use of punishment
(c) pairing (d) stimuli

91; (a) avoidance conditioning (b) warning signal
(c) escape training (d) specific response

92. (a) avoidance conditioning (b) warning signal
(c) escape training (d) specific response

93. (a) avoidance (b) specific response
(c) active (d) passive

94. (a) avoidance (b) specific response
(c) active (d) passive

Now look at the following passage and fill in the blanks with the correct word from the list below.

Geology is vitally important for the needs and industries of mankind, (95) , thousands of geologists are actively engaged in locating and exploring the mineral resources of the earth, (96) coal and iron. (97), geologists are (98) directly concerned with the study of water supply. (99), many engineering projects, (100) tunnels, canals docks and reservoirs call for geological advice in the selection of sites and materials.

95. (a) However (b) For example
(c) In addition (d) such as

96. (a) also (b) such as
(c) moreover (d) indeed

97. (a) In addition (b) However
(c) For example (d) On the other hand

98. (a) however (b) for example
(c) in addition (d) also

99. (a) At last (b) Finally
(c) On the contrary (d) Similarly

100. (a) unlike (b) indeed
(c) also (d) for example

UNIVERSITI SAINS MALAYSIA
PUSAT BAHASA DAN TERJEMAHAN

ENGLISH LANGUAGE PLACEMENT TEST
FORM D

PART 3 - 1½ hours.

Part 3 is a test of your English comprehension. There are five passages, each with eight questions - a total of 40 questions.

Read through each passage in turn and then answer the questions on it.

Mark your answers - (a), (b), (c) or (d) - in thick pencil on the answer sheet.

Now turn over and begin Part 3. Work as quickly as possible.

...2/-

PASSAGE A

The Experimenter Effect

It is significant that a problem which perplexed some of the most influential scientists of Germany in 1904 was resolved at that time, yet should contaminate psychology investigations of the present day, that is, the experimenter's influence on his subjects. The amazing horse of Mr. von Osten caused an uproar throughout all of Germany which Professor Stumpf and his co-workers, through meticulous investigation, demonstrated to be the result of the questioners' unintentional, involuntary cues utilized by the animal. This incident dramatically emphasized the stimulus value of "unconscious" cues emitted by an experimenter to his animal subjects. Even though questioners of "Hans" were aware that this might be the explanation for his feats and were most careful in attempting to refrain from allowing him this advantage, the unconscious cues were still emitted until the situation was carefully analyzed and the specific variables controlled (Pfungst, 1911).

McGuigan (1963) states:

While we have traditionally recognized that the characteristics of an experimenter may indeed influence behavior, it is important to observe that we have not seriously, attempted to study him as an independent variable (p.421);

However, Stumpf with his careful, detailed measurements of questioners' cues began the study of the experimenter as an independent variable in 1904, but not until recently has this problem been considered by experimental psychologists for study (Cordaro and Ison, 1963; McGuigan, 1963; Rosenthal and Halas, 1962). Clinical psychologists have long led the way in this aspect of investigation. The personal effect of examiners upon patients' performance in clinical tests was initiated as an object of study 35 years ago (Marine, 1929). Yet, psychologists working in the laboratory have not been completely unaware of the implications of experimenter influence upon subjects.

Ebbinghaus (1913) in discussing the effects of early data returns upon psychological research stated:

It is unavoidable that, after the observation of the numerical results, suppositions should arise as to general principles which are concealed in them and which occasionally give hints as to their presence. As the investigations are carried further, these suppositions, as well as those present at the beginning, constitute a complicating factor which probably has a definite influence upon the subsequent results (pp. 28-29).

Pavlov, noting the apprent increase in learning ability of successive generations of mice in experiments on the inheritance of acquired characteristics, suggested that an increase in the teaching ability of experimenters may have, in fact, constituted the critical variable (Gruenberg, 1929, p.327).

The foregoing yields some indication of the scope inherent in this phenomenon.

...3/-

101. The main idea of this passage is that

(a) a problem first stated in Germany in 1904 remains unsolved today

(b) animals should only be used in experiments if they are unconscious

(c) experimenters in psychology often unconsciously influence their subjects

(d) experiments should be considered as independent variables.

102. "Hans" (line 10) was

(a) a psychologist

(b) an experimenter

(c) a German

(d) a horse

103 The problem of the experimenter as an independent variable

(a) is a matter for detailed measurement

(b) has not recently been considered by experimental psychologists

(c) has long been studied by clinical psychologists

(d) has never been clear to psychologists working in the laboratory.

104. Ebbinghaus (1913) highlighted the problem of

(a) the observation of numerical results

(b) the general principles concealed in observations.

(c) complicating factors in experiments

(d) suppositions as a complicating factor

105. Pavlov suggested that successive generations of mice showed an increase in learning ability because

(a) experimenters got better at teaching them

(b) the mice learned from their mistakes

(c) learning characteristics are inherited

(d) the experimenters became less critical

...b/...

106. From this passage we may conclude that

(a) psychologists should not carry out experiments

(b) the subjects of experiments should be carefully controlled

(c) experimenters may have an effect upon the results of an experiment.

(d) animals are more intelligent than we suppose.

107. "to refrain" (line 11) means

(a) to hold oneself back

(b) to stop oneself

(c) to permit oneself

(d) to indulge oneself

108. "this" (line 10) refers to

(a) the animal used in the experiment

(b) the stimulus value of "unconscious" cues

(c) the explanation for his feats

(d) refraining from allowing him this advantange

...5/-

PASSAGE B

Some problems of automation

In an automated plant there is frequently very little for the operator to do; the rooms are usually kept at comfortable temperatures and the noise levels are low. It is a common experience that as the environment becomes more comfortable and stimulation is reduced, so men become drowsy or bored and inattentive. This condition reduces efficiency in the sense that quick and effective responses to emergencies suffer, and it also means that danger symptoms are often not spotted until it is too late. A number of techniques for overcoming these problems are available, and active research is going on, for example, into methods of improving the efficiency of signal detection, that is, the ability to pick out an important signal from other, less important or irrelevant signals.

One widely used method is the false alarm. Here artificial fault conditions are signalled to the operator who does not know at the time whether there is a real emergency or not and he must take the appropriate action as if it were a real crisis. This cannot be used in certain plants without the operator immediately being aware that it is a false alarm, and in any case too many false alarms build up a negative attitude in the operator. A certain number of test alarms can be useful but they must be very carefully planned so that they are indistinguishable from the real thing and are relatively unpredictable. Experimental studies of men doing watch-keeping tasks have suggested a number of other methods for improving alertness: for example, a certain amount of noise or background music and variations in temperature and humidity are useful. Much more attention could be paid to making the environment in control rooms more stimulating without distracting the man from his primary task. One important factor often overlooked is the beneficial effects of social contact with other people - even telephone contact is valuable. It may, for example, be worthwhile using a man to deliver a message which could well be done by an electronic link, since letting the man do the task enables him to meet other people.

The problems of ensuring appropriate actions in an emergency is in many ways more difficult to solve. Many fault conditions can be anticipated and suitable emergency drills prepared; but the very nature of the modern complex plant means that it is virtually impossible, to predict all the different things which can go wrong. It is still necessary to rely on the operator recognising the presence of danger conditions and taking the appropriate actions. This means that operators may have to have a much more detailed knowledge of the plant and how it works than may be apparent at first sight.

(From "Ergonomics and Automation" by R.J. Beishon)

...6/-

109. In order to overcome the problems which result from comfortable working condition, workers are given

(a) higher salaries

(b) irrelevant signals

(c) test alarms

(d) watch-keeping tasks

110. If there are a lot of false alarms, the worker

(a) does not know whether there is a real emergency or not

(b) is made much more aware and conscious of danger

(c) grows relatively upredictable

(d) becomes used to the alarms and does not respond properly.

111. According to the writer, the temperature and the humidity in an automated plant should be

(a) varied

(b) raised

(c) reduced

(d) kept constant

112. Social contact between workers is often

(a) valuable

(b) harmful

(c) neither valuable nor harmful

(d) valuable for the people concerned but harmful for efficiency

113. Because modern factories are so complex,

(a) the operator cannot possibly know when something is wrong with a machine

(b) no one can say with certainty when anything will go wrong

(c) the operators cannot have a very detailed knowledge of the plant

(d) emergency drills have to be prepared and anticipated.

114. "drowsy" (line 5) means

(a) sleepy (b) lethargic

(c) dull (d) distracted

115. "it" (line 16) refers to

(a) a real emergency (b) the appropriate action

(c) the false alarm (d) the time

116. "it" (line 30) refers to

(a) telephone contact (b) using a man to deliver a message

(c) using an electronic link (d) one important factor

.../7-

PASSAGE C

Modern surgery

The need for a surgical operation, especially an emergency operation, almost always comes as a severe shock to the patient and his family. Despite modern advances, most people still have an irrational fear of hospitals and anaesthetics. Patients do not often believe they really need surgery - cutting into a part of the body as opposed to treatment with drugs.

In the early years of this century there was little specialization in surgery. A good surgeon was capable of performing almost every operation that had been devised up to that time. Today the situation is different. Operations are now being carried out that were not even dreamed of fifty years ago. The heart can be safely opened and its valves repaired. Clogged blood vessels can be cleaned out, and broken ones mended or replaced. A lung, the whole stomach, or even part of the brain can be removed and still permit the patient to live a comfortable and satisfactory life. However, not every surgeon wants to, or is qualified to carry out every type of modern operation.

The scope of surgery has increased remarkably in this century. Its safety has increased too. Deaths from most operations are about 20% of what they were in 1910 and surgery has been extended in many directions, for example to certain types of birth defects in newborn babies, and, at the other end of the scale, to life-saving operations for the octogenarian. The hospital stay after surgery has been shortened to as little as a week for most major operations. Most patients are out of bed on the day after an operation and may be back at work in two or three weeks.

Many developments in modern surgery are almost incredible. They include the replacement of damaged blood vessels with simulated ones made of plastic; the replacement of heart valves with plastic substitutes; the transplanting of tissues such as the lens of the eye; the invention of the artificial kidney to clean the blood of poisons at regular intervals and the development of heart and lung machines to keep patients alive during very long operations. All these things open a hopeful vista for the future of surgery.

One of the most revolutionary areas of modern surgery is that of organ transplants. Until a few years ago, no person, except an identical twin, was able to accept into his body the tissues of another person without reacting against them and eventually killing them. Recently, however, it has been discovered that with the use of x-rays and special drugs, it is possible to graft tissues from one person to another which will survive for periods of a year or more. Kidneys have been successfully transplanted between non-identical twins. Heart and lung transplants have been reasonably successful in animals, though rejection problems in humans have yet to be solved.

'Spare parts' surgery, the simple routine replacement of all worn-out organs by new ones, is still a dream of the distant future. As yet, surgery is not ready for such miracles. In the meantime, you can be happy if your doctor says to you, 'Yes, I think it is possible to operate on you for this condition.'

...8.-

117. Most people are afraid of being operated on

(a) in spite of improvements in modern surgery
(b) because they think modern drugs are dangerous
(c) because they do not believe they need anaesthetics
(d) unless it is an emergency operation

118. Compared with modern surgeons, those in the early years of the century

(a) had less to learn about surgery
(b) needed more knowledge
(c) could perform every operation known today
(d) were move trusted by their patients

119. Modern surgeons

(a) do not like to perform operations of the new type
(b) are not as highly qualified as the older ones
(c) are obliged to specialize more than their predecessors
(d) often perform operations which are not really needed.

120. Some of the more astonishing innovations in modern surgery include

(a) ear, nose and throat transplants
(b) valveless plastic hearts
(c) plastic heart valves
(d) leg transplants

121. The main difficulty with organ transplants is

(a) it is difficult to find organs of exactly the same size
(b) only identical twins can accept into their bodies the tissues of another
(c) the body's tendency to reject alien tissues
(d) the patient is not allowed to used drugs after them

...9/-

122. An octogenarian (line 20) is

(a) an eighteen-year old

(b) a person in his eighties

(c) a patient having his eighth operation

(d) someone born in the 1980's

123. "Them" (line 35) refers to

(a) identical twins

(b) the tissues of another person

(c) the people receiving a transplant

(d) the people giving an organ for transplant

124. "it" (line 35) refers to

(a) the possibility of grafting tissues from are person to another

(b) the use of x-rays and special drugs

(c) transplanting organs from one body into another

(d) ability to accept the tissues of another person.

PASSAGE D

Before we consider what mechanisms could possibly underly the metal-bending effect, we must first help ourselves by starting with a brief summary of what it is that we shall have to try to explain. The multitude of the accounts makes it clear that we are dealing with a genuine effect which can happen sometimes as a result of direct contact with a subject, and sometimes without it. The main action in the case of direct contact appears to be that of gentle stroking by the fingers of one hand. The length of time taken to cause an appreciable bend seems to vary, but it is normally less than thirty minutes and more than two or three; moreover, for a particular subject it can vary considerably from one day to the next.

So also the attitude of the subject during the stroking process varies considerably. Some concentrate very hard and focus on the object. Others, again, may take little direct notice of the piéce of cutlery they are gently stroking.

One curious feature of the bending process is that it appears to go in brief steps; a spoon or fork can bend through many degrees in a fraction of a second. This often happens when the observer's attention has shifted from the object he is trying to bend. Indeed this feature of bending not happening when the object is being watched is very common. It seems to be correlated with the presence of sceptics or others who have a poor relationship with the subject. Another feature of the metal bending process is that it appears to require a large amount of energy; feelings of fatigue usually are experienced at the end of a bending session.

The exhaustion which follows metal-bending can be regarded as possible evidence for the emission of energy during the bending process. No child has been weighed before and after a bending session, so as yet there are no hard data on this. But it does show that there is an effect here which warrants further investigation. Energy emission by the subject would also be consistent with one of the fundamental principles of physics, energy conservation - the idea that energy can only be gained at the expense of something else.

...11/-

125. People who have metal-bending powers

(a) have to try to explain what they doing

(b) take much longer to cause bends on some days than others

(c) gently stroke the fingers of their hands together

(d) practise between 2 minutes and 30 minutes a day

126. In the passage, how many attitudes of the subject are mentioned?

(a) 1 (b) 2 (c) 3 (d) 4

127. When the observer is not concentrating an the object

(a) a spoon or fork can bend appreciably in a fraction of a second

(b) it takes a longer time to bend a spoon or fork

(c) he feels his energy going out of him

(d) he uses the time to increase his energy

128. Feelings of fatigue are usually experienced at the end of a bending bending session because

(a) it has usually lasted a long time

(b) an extra effort has to made in the presence of sceptics

(c) it requires a large amount of energy

(d) the emotional relationships between subject and objects are difficult to establish

129. Further investigation is needed into

(a) the weight of children before and after a bending session

(b) the hard data of metal bending

(c) the process by which a person bends spoons and forks

(d) the exhaustion which follows metal-bending

130. "it" (line 2) refers to

(a) to help ourselves

(b) considering the mechanisms of metal-bending

(c) making a brief summary

(d) what we shall have to try to explain

131. "it" (line 19) refers to

(a) when the observer's attention has shifted

(b) the presence of sceptics or others.....

(c) bending not happening when the object is being watched

(d) to be correlated

132. "shifted" (line 17) means

(a) concentrated

(b) moved

(c) described

(d) bent

...12/-

PASSAGE E

Since the idea of intermediate technology was first put forward, a number of objections have been raised. The most immediate objections are psychological: 'You are trying to withhold the best and make us put up with something inferior and out-dated. This is the voice of those who are not in need, who can help themselves and want to be assisted in reaching a higher standard of living at once. It is not the voice of those with whom we are here concerned, the poverty-stricken multitudes who lack any real basis of existence, whether in rural or in urban areas, who have neither 'the best' nor 'the second best' but go short of even the most essential means of subsistence.

There are economists who believe that development policy can be derived from certain allegedly fixed ratios, such as the capital/output ratio. Their argument runs as follows: The amount of available capital is given. Now, you may concentrate it on a small number of highly capitalized workplaces, or you may spread it thinly over a large number of cheap workplaces. If you do the latter, you obtain less total output than if you do the former, you therefore fail to achieve the quickest possible rate of economic growth. Not only capital but also 'wages goods' are held to be a given quantity, and this quantity determines 'the limits on wages employment in any country at any given time'.

The first thing that might be said about these arguments is that they are evidently static in character and fail to take account of the dynamics of development. To do justice to the real situation it is necessary to consider the reactions and capabilities of people, and not confine oneself to machinery or abstract concepts. As we have seen before, it is wrong to assume that the most sophisticated equipment, transplanted into an unsophisticated environment, will be regularly worked at full capacity, and if capacity utilization is low, then the capital/output ratio is also low. It is therefore fallacious to treat capital/output ratios as technological facts, when they are so largely dependent on quite other factors.

The question must be asked moreover whether there is such a law, as Dr Kaldor asserts, that the capital-output ratio grows if capital is concentrated on fewer workplaces. No one with the alightest industrial experience would ever claim to have noticed the existence of such a 'law', nor is there any foundation for it in any science. Mechanization and automation are introduced to increase the productivity of labour, i.e. the worker/output ratio, and their effect on the capital/output ratio may just as well be negative as it may be positive. Countless examples can be quoted where advances in technology eliminate workplaces at the cost of an additional input of capital without affecting the volume of output. It is therefore quite untrue to assert that a given amount of capital invariably and necessarily produces the biggest total output when it is concentrated on the smallest number of workplaces.

The greatest weakness of the argument however lies in taking 'capital' - and even 'wages goods' - as 'given quantities' in an underemployed economy. Here again, the static outlook inevitably leads to erroneous conclusions. The output of even a poorly equipped man can be a positive contribution, to 'capital' as well as to 'wages goods'. The distinction between those two is by no means as definite as the econometricians are inclined to think, because the definition of 'capital' itself depends decisively on the level of technology employed.

...13/-

133. Objections to the idea of intermediate technology

(a) have a firm psychological basis

(b) ignore the needs of those really in need

(c) are generally considered out-dated

(d) are raised by those who live in urban areas

134. From paragraph 2 we may conclude that the author believes that

(a) capital and wages goods are a fixed quantity

(b) a large number of cheap work places can be financed for a small number of highly capitalized ones

(c) the capital/output ratio is not fixed

(d) the quickest rate of economic growth will be achieved by concentrating on highly capitalized workplaces

135. Concentrating only an machinery or abstract concepts

(a) fails to take account of the dynamics of development

(b) is necessary to do justice to the real situation

(c) assumes that sophisticated equipment cannot be transplanted into an unsophisticated environment

(d) ensures that equipment will be worked at full capacity

136. The worker/output ratio and the capital/output ratio

(a) form the basis of Dr Kaldar's "law"

(b) are ignored by those with industrial experience.

(c) do not have any foundation in any science

(d) are not necessarily connected

137. The "static outlook" of the last paragraph

(a) helps to explain "capital" and "wages goods" as "given quantities"

(b) describes the normal situation in an underemployed economy

(c) causes us srongly to conclude that a poorly equipped man's output makes no positive contribution

(d) depends on the definition of 'capital' being decisively dependent on the level of technology employed

...14/-

138. "allegedly" (line 11) means

(a) permanently
(b) constantly
(c) unlikely
(d) supposedly

139. fallacious (line 26) means

(a) wrongly concluded
(b) falsely argued
(c) properly considered
(d) rightly disproved

140. "it" (line 38) refers to

(a) to assert that .... workplaces
(b) to quote examples where ----- output
(c) the volume of output
(d) a given amount of capital

UNIVERSITI SAINS MALAYSIA
PUSAT BAHASA DAN TERJEMAHAN

ENGLISH LANGUAGE PLACEMENT TEST
FORM D

PART 4 - ½ hour

Part 4 is a test of your ability to summarize a passage of English.

Read quickly through the passage on the next page, then write a summary of the passage in English.

You should use between 80 - 150 words.

Use the separate answer sheet for your answer and for your rough work.

THE RESISTANCE OF INSECT PESTS TO INSECTICIDES

The ultimate type of resistance is that in which the insect changes its normal physiology so that it is no longer sensitive to the insecticide. A change of this kind seems to be the explanation of the type of resistance involving a large number of chlorinated compounds like dieldrin. The mode of action of these compounds, however, is quite obscure, so that at present it is scarcely possible to discover how insects become immune to them.

Research in the past fifteen years has revealed a great deal about the nature of resistance, but in no single case have we been able to overcome it completely. In other words, when resistance has developed to a particular insecticide, no means have been found to restore permanently the former effectiveness of that insecticide.

Considering the present situation, it may cause surprise, in view of the large number of reports of resistance from so many important specides all over the world, that the impact on insect control programmes is not more drastic. There are two reasons for this. Firstly, many instances of resistance are more or less localized. For example, dieldrin resistance in the major African malaria vector, Anopheles gambiae, is confined to the west of Africa, though the mosquito occurs in East and South Africa and is equally attacked by insecticides in those regions. One may begin to hope that the genetical potential for developing resistance is lacking in some natural populations of pest insects. Secondly, only a limited number of species show resistance to the two groups of chlorinated insecticides. Until this double resistance develops, it is possible to use either one or the other of these two classes of insecticide and still maintain effective control. Unfortunately, however, the instances of double resistance are growing. By 1960, twenty species of public health importance had developed resistance to both groups of chiorinated insecticides. In addition, four species had developed resistance to organo-phosphorus compounds as well - in other words, treble resistance. It must, then, be concluded that resistance is likely to become a more severe problem in the future then it is at present.

Naturally, a great deal of thought has been given to possible ways of preventing the emergence of resistance. One suggestion has been the use of mixtures of two different types of insecticide, with the idea that one of them should eliminate the individuals resistant to the other. This principle has been found useful in preventing resistance to antibiotics in bacteria. Unfortunately, the few practical trials have not been encouraging, for the mixtures have merely developed a double resistance to the two insecticides employed.

In brief, there is as yet no known way of obtaining the benefits of the new insecticides without some risk of provoking resistance. For this reason, it would seem unwise to use insecticides regularly, on a very large scale, unless there is some vital object to be attained. In such cases, the use of insecticides should be combined with other measures, for examples (as regards insect-borne disease) strong efforts to improve general hygiene.

BUSVINE, J. R., 'The Challenge of Insecticide Resistance', Penguin Science Survey B, 1966, Penguin Books Ltd.

UNIVERSITI SAINS MALAYSIA
PUSAT BAHASA DAN TERJEMAHAN

ENGLISH LANGUAGE PLACEMENT TEST

FORM A

Part 3 - 1½ hours

Part 3 is a test of your English comprehension. There are five passages, each with eight questions - a total of 40 questions.

Read through each passage in turn and then answer the questions on it.

Mark your answers - (a), (b), (c) or (d) - in thick pencil on the answer sheet.

Now turn over and begin Part 3. Work as quickly as possible.

Passage D

Can Other Animals Acquire Language?

Animals other than humans have not developed communications comparable to human language. But is it possible that other animals have the capacity to learn a language if they are adequately taught? Obviously, this is a fascinating notion. The idea of communicating directly with another species has long been a part of human folklore and children's fantasies. But on a scientific level, the question of whether animals can learn a language is important primarily because it relates to the controversy between the cognitive and the learning approaches to language. If language is dependent on and is actually an outgrown of the intellectual structure of the human mind, there is the strong supposition that only humans are capable of using language. Therefore, Noam Chomsky and other psycholinguists have argued that only humans can learn a language, while most behaviour feel that with sufficient patience it should be possible to teach an animal some sort of language. Although the two schools of thought clearly differ on this point, it is not really a crucial test of the two theories. If a chimpanzee can master a simple language all it would mean is that the chimp's intellectual capacity and brain structure are more similar to ours than we thought. It would not necessarily imply that our intellectual structure is unimportant in our own mastery of language. Thus, teaching an animal language is an impressive demonstration of the power of learning techniques, but it is not evidence that language is developed entirely through learning.

On the other hand, the question of whether other animals can learn a language is fascinating in its own right, aside from its value as a test of the two theories of language development. Accordingly, whatever one's position on the theoretical dispute, we must consider training an animal to use language a dramatic accomplishment.

125. The capacity to learn a language

a. is unique to humans
b. is common to humans and animals
c. may be present in animals, but we can't be sure
d. can be seen in animals if they are adequately taught

126. Communicating directly with other species

a. has long been a feature of animal behaviour
b. has long been a feature of human stories
c. is only thought of in children's fantasies
d. has little to do with scientific research

127. The cognitive approach to language implies that

a. language is uniquely human
b. language can be learned both by animals and humans
c. animal minds have no intellectual structure
d. chimpanzees can master a simple language

128. The learning approach to language implies that

a. language is dependent on the human mind
b. it should be possible to teach an animal some sort of language
c. with patience anyone can learn anything
d. school is an essential place for language learning.

129. If a chimpanzee can master a simple language, this would mean that

a. the learning approach was proved right
b. the cognitive approach was proved right
c. either the cognitive or the learning approach was proved right
d. neither the cognitive nor the learning approach was proved right.

130. Training an animal to use language would be

a. a good test of learning theory
b. a dramatic achievement
c. interesting but not scientific
d. an intellectual accomplishment

131. "adequately" (line 3) means

a. simply
b. well
c. satisfactorily
d. scientifically

132. "it" (line 13) refers to

a. sufficient patience
b. to teach an animal some sort of language
c. the capability of using language
d. the intellectual structure of the human mind.

Passage E

The 'weak' view about the technology of communications - that it leads to certain opportunities which may or may not be taken up - is in itself much less exciting and challenging than the stronger, causal view. It is also much less tidy and more difficult to generalize, since social choices and additional influences also enter in, so that likely results become near-impossible to predict and the range of alternative possibilities is almost infinite. Nevertheless it does accord much better with the detailed empirical evidence.

This view is one that is in any case more likely to commend itself to many social scientists. This is because the significance of, say, political and social factors as motivating forces in their own right can be recognized, rather than regarded as primarily conditioned by technological factors.

Certainly the picture of opportunities being provided in certain respects by various communications technologies seems to fit well with the detailed evidence on social and economic developments. The medium in itself cannot give rise to social consequences - it must be used. The mere technical existence of writing cannot affect social change: it is its use, who uses it, who controls it, what it is used for, how it fits into the power structure, how widely it is distributed - all these social and political factors radically affect the possible consequences. Thus the implications of writing are very different when, for instance, it is strictly confined to priests and rulers and largely concerned with religion (as in early Egypt) from the situation where there is widespread literacy. And this is different yet again from the situation in a contemporary developing country when adult illiteracy may be increasing in absolute numbers, and writing is used for a whole range of purposes, but literacy is largely confined to an elite of relatively young people who as a result take on the best paid and most powerful jobs - with political and economic consequences that can be imagined.

This is not to throw away the case for emphasizing the technology of communications completely - it is only to show that it is a more complex situation than at first envisaged. Clearly the various technologies of communication do provide opportunities, and, conversely, their absence provides constraints. Without writing extensive and accurate communication over time and space is impossible: and it is essential to bear this in mind in analysing non-literate societies. Similarly it is only with telecommunications that instantaneous communication over distance is possible and that opportunities provided by this fact can be exploited. The very important constraints and opportunities provided by media are forced to our attention by even the weak form of the theory about communications - technology. And, despite its untidiness and lack of clear definition as well as its slightly tame impression compared to the exciting extravagances of the 'strong' case, it still seems one that should be considered extremely seriously by social scientists as both providing a 'model' for illuminating reality and leading to further research as well as fitting in with the facts as known at present.

Passage E

133. The "strong" view about the technology of communications is likely to be that

a. communications technology is the single determining cause of social development.

b. communications technology inevitably leads to industrial progress.

c. economic development is determined by communications technology.

d. social developments can be predicted from an analysis of communications technology.

134. Social scientists are likely

a. to hold the "strong" view about the technology of communications.

b. to be motivated by political and social factors.

c. to be regarded as primarily conditioned by technological factors.

d. to believe that the technologyof communications leads to certain opportunities which may not be taken up.

135. The example of the technology of writing (in paragraph 3) shows that

a. adult literacy is increasing in absolute numbers in developing countries.

b. literacy is largely confirmed to an elite of relatively young people.

c. the medium of communication itself cannot give rise to social consequences.

d. social and political factors do not necessarily have an effect on communications technologies.

136. The writer believes that

a. the political and economic consequences of literacy must be imagined.

b. the technology of communications is still worth emphasizing.

c. we should abandon the idea of a technology of communications.

d. adult illiteracy is not necessarily a bad thing.

137. The writer conludes that

a. neither the "strong" nor the "weak" views are likely to be useful.

b. the technology of communications results in a more complex situation that at first envisaged.

c. the various technologies of communication do provide opportunities.

d. the "weak" view fits in with the facts as known at present.

138. "accord" (line 7) means

a. agree

b. play

c. argue

d. demonstrate

139. "envisaged" (line 32) means

a. described

b. imagined

c. hypothesized

d. confronted

140. "it" (line 36) refers to

a. only with telecommunications

b. to bear this in mind in analysing non-literate societies.

c. that instantaneous communication over distance is possible.

d. extensive and accurate communication over time and space.

UNIVERSITI SAINS MALAYSIA
PUSAT BAHASA DAN TERJEMAHAN

ENGLISH LANGUAGE PLACEMENT TEST

FORM A

Part 4 - ½ hour

Part 4 is a test of your ability to summarize a passage of English.

Read quickly through the passage on the next page, then write a summary of the passage in English.

You should use between 80 - 150 words.

Use the separate answer sheet for your answer and for your rough work.

The period of evolution into which man has brought himself by his own efforts can best be described as 'social evolution', rather than 'biological evolution'. The term 'social evolution' is not entirely apt, because the changes in the achievements of the species depend predominantly on the intellectual abilities of mankind arising out of the structure of the human brain and the use which men make of the products of their brains. It will serve because the evolutionary effects are only brought to bear by making use of social organisations. Our ability to fly comes from the development of aeronautical sciences. Nevertheless, aeroplanes can only be made by nations organized as industrial communities. The profound changes in Western society due to the increasing application of automation similarly depend on the social organization. In the new world of automation when we shall be economically rich - almost as if we had all won prizes in the football pools - and shall need to spend only a small proportion of our time in factories and offices, the whole business will depend on the availability of large amounts of capital. This collected wealth is a social phenomenon.

But although the social evolution which moves so quickly that we can actually see it happening is due to brains rather than biology, it is evolution nevertheless. And the basis of evolution is the survival of the fittest. This raises a knotty point. Let me quote what the late Professor Joad had to say about it:

> Why does evolution go on, and go on to complicate our structure so unnecessarily that, instead of becoming more fitted to our physical environment than we used to be, we are less? A degree of adaptation which, from the purely physical point of view, would put the average human being to shame has been achieved by living organisms thousands of years ago. The inference is irresistible, that the achievement by life of mere adaptation is not enough, but that living beings are evolved at more complicated and therefore more dangerous levels, in the endeavour to attain higher forms of life. The amoeba, in short, is superseded by man, not because man is better-adapted, life, but because he is better-quality life.

As soon as we talk about 'better quality' in the context of the scientific facts of evolution we introduce a new factor with which science does not generally deal. This is the implication of value, that one kind of creature is better than another - in short, that one kind of life, with all its attributes of behaviour and intellectual ability, is at a higher plane, in whatever standards of good and bad we may set, than another. The sciences, neither physics, chemistry, nor, indeed, biology either, have anything to say about moral values. The tenet of the 'survival of the fittest' has brought living creatures through the principles

of evolution from the simplest type of animate life, through the various stages of zoological progression up to the earliest, low-browed man fashioning stone implements in a cave. The life of such men has been described as 'nasty, brutish, and short'. We have come a long way since then. In a short thousand years well-to-do Europeans have progressed from the draughty castle to the civilized gentlemean's drawing-room. Today, the cumulative advance in scientific knowledge has brought us to the edge of the greatest forward achievement of all.

Magnus Pyke

UNIVERSITI SAINS MALAYSIA
PUSAT BAHASA DAN TERJEMAHAN

ENGLISH LANGUAGE PLACEMENT TEST

FORM B

PART 3 - 1½ hours

Part 3 is a test of your English comprehension. There are five passages, each with eight questions - a total of 40 questions.

Read through each passage in turn and then answer the questions on it.

Mark your answers - (a), (b), (c), or (d) - in thick pencil on the answer sheet.

Now turn over and begin Part 3. Work as quickly as possible.

Passage C

Why did they butcher it so?

The question why comes back again and again. Why did they butcher it so? These were not people running away from technology. These were the technologists themselves. They sat down to do a job and they performed it like chimpanzees. Nothing personal in it. There was no obvious reason for it. And I tried to think back into that shop, that nightmare place to try to remember anything that could have been the cause.

The radio was a clue. You can't really think hard about what you're doing and listen to the radio at the same time. Maybe they didn't see their job as having anything to do with hard thought, just wrench twiddling. If you can twiddle wrenches while listening to the radio, that's more enjoyable.

Their speed was another clue. They were really slopping things around in a hurry and not looking where they slopped them. More money that way - if you don't stop to think that it usually takes longer or comes out worse.

But the biggest clue seemed to be their expressions. They were hard to explain. Good-natured, friendly, easygoing - and uninvolved. There were like spectators. You had the feeling they had just wandered in there themselves and somebody had handed them a wrench. There was no identification with the job. No saying, "I am a mechanic." At 5 p.m. or whenever their eight hours were in, you knew they would cut it off and not have another thought about their work. They were already trying not to have any thoughts about their work on the job. In their own way they were achieving the same thing my companions were, living with technology without really having anything to do with it. Or rather, they had something to do with it, but their own selves were outside of it, detached, removed. They were involved in it but not in such a way as to care.

While at work I was thinking about this same lack of care in the digital computer manuals I was editing. Writing and editing technical manuals is what I do for a living the other eleven months of the year and I knew they were full of errors, ambiguities, omissions and information so completely screwed up you had to read them six times to make any sense out of them. But what struck me for the first time was the agreement of these manuals with the spectator attitude I had seen in the shop. These were spectator manuals. It was built into the format of them. Implicit in every line is the idea that "There is the machine, isolated in time and in space from everything else in the universe. It has no relationship to you, you have no relationship to it, other than to turn certain switches, maintain voltage levels, check for error conditions ..." and so on. That's it. The mechanics in their attitude toward the machine were really taking no different an attitude from the manual's toward the machine, or from the attitude I had when I brought it in there. We were all spectators. And it occured to me there is no manual that deals with the real business of motorcycle maintenance, the most important aspect of all. Caring about what you are doing is considered either unimportant or taken for granted.

(From "Zen and the Art of Motorcycle Maintenance"
By Robert M. Pirsig)

Passage C

117. "The radio was a clue" (line 8). The writer is referring to the reason why

(a) the mechanics butchered the motorcycle

(b) the writer disliked the place

(c) the mechanics imitated chimpanzees

(d) the writer scorned modern technology

118. The mechanics did their jobs

(a) without bothering to explain anything to their fellow workers.

(b) in a very friendly way

(c) quickly and without experiencing and difficulty

(d) objectively and without any real interest

119. From the passage we can gather that the mechanics worked

(a) five hours a day

(b) eight hours a day

(c) thirteen hours a day

(d) sixteen hours a day

120. The writer complained that in technical manuals machines were treated as it

(a) they were completely unconnected to the people using them

(b) they had a life and individuality of their own.

(c) they were usually made to appear very attractive to people

(d) they dominated everything in the universe.

121. The writer's main argument is that

(a) modern technology is beginning to dominate man.

(b) people nowadays prefer to be passive rather than active.

(c) no one cares about what he is doing.

(d) many workers will always be dissatisfied with their work.

122. "wrench" (line 21) is

(a) an instrument used in designing motorcycle engines

(b) a kind of dance which a person can perform on his own.

(c) a tool for tightening or loosening bolts.

(d) a small but important part of an engine.

123. "screwed up" (line 34) means

(a) confused

(b) technical

(c) detailed

(d) inaccurate

124. "it" (line 29) refers to

(a) the work they were doing

(b) the same thing

(c) living

(d) technology

Passage D

When we are in a certain intellectual mood, we seem to find clashes between the things that scientists tell us about our furniture, clothes and limbs, and the things that we tell about them. We are apt to express these felt rivalries by saying that the world whose parts and members are described by scientists is different from the world whose parts and members we describe ourselves, and yet, since there can be only one world, one of these seeming worlds must be a dummy world. Moreover, as no one nowadays is hardy enough to say 'Bo' to science, it must be the world that we ourselves describe which is the dummy-world.

As a preface to the serious part of the argument I want to deflate two over-inflated ideas, from which derives not the cogency but some of the persuasivences of the argument for the irreconcilability of the world of science with the everyday world. One is the idea of science, the other that of world.

(a) There is no such animal as 'Science'. There are scores of sciences. Most of these sciences are such that acquaintance-ship with them or, what is even more captivating hearsay knowledge about them has not the slightest tendency to make us contrast their world with the everyday world. Philology is a science, but not even popularizations of its discoveries would make anyone feel that the world of philology cannot be accom-modated by the world of familiar people, things and happenings. Let philologists discover everything discoverable about the structures and origins of the expressions that we use; yet their discoveries have no tendency to make us write off as mere dummies the expressions that we use and that philologists also use. The sole dividedness of mind that is induced in us by learn-ing any of the lessons of philology is akin to that which we sometimes experience when told, say, that our old, familiar paper-weight was once an axe-head used by a prehistoric warrior. Something utterly ordinary becomes also, just for the moment, charged with history. A mere paper-weight becomes also, just for the moment, a death-dealing weapon. But that is all.

Nor do most of the other sciences give us the feeling that we live our daily lives in a bubble-world. Botanists, entomologists, meteorologists, and geologists do not seem to threaten the walls, floors and ceilings of our common dwelling-place. On the con-trary, they seem to increase the quantity and improve the arrangement of its furniture. Nor even, as might be supposed, do all branches of physical science engender in us the idea that our everyday world is a dummy-world. The discoveries and theories of astronomers and astro-physicists may make us feel that the earth is very small, but only by making us feel that the heavens are very big. The gnawing suspicion that both the terrestrial and the super-terrestrial alike are merely painted stage- canvas is not begotten by even hearsay knowledge of the physics of the immense. It is not begotten, either, by hearsay knowledge of the physics of the middle-sized. The theory of the pendulum, the cannon-ball, the water-pump, the fulcrum, the balloon and the steam-engine does not by itself drive us to vote

between the everyday world and the so-called world of science. Even the comparatively minute can be accommodated by us without theoretical heart-searchings in our everyday world. Pollen-grains, frost-crystals and bacteria, though revealed only through the microscope, do not by themselves make us doubt whether middle-sized and immense things may not belong.
60 We always knew that there were things too small to be seen with the naked eye; the magnifying-glass and the microscope have surprised us not by establishing their existence but by disclosing their variety and, in some cases, their importance.

Gilbert Ryle

Passage D

125. The implication of paragraph 1 is that

(a) scientists tell us things about the everyday world that we already know.

(b) scientists describe the real world while we describe the dummy world.

(c) our experience of the world is different from other kinds of experience

(d) science has shaped our experience of the world we know.

126. The example of philology in paragraph 3 shows that

(a) science is not one but many

(b) the sciences do not force us to think of the familiar world as unreal.

(c) popularization can be a dangerous thing in the science.

(d) the contrast between the sciences and the world we experience is very great.

127. The example of astronomers and astrophysicists in paragraph 4 show that

(a) most sciences do not threaten the familiar world

(b) our everyday world is a bubble world

(c) there are exceptions even in the physical sciences to the idea being suggested

(d) the earth is very small and the heavens are very big.

128. The physics of the middle-sized

(a) drives us to choose between the everyday world and the world of science.

(b) differs in this context from the physics of the immense.

(c) helps explain how the pendulum and the cannon ball work.

(d) is not responsible for our feeling that ours is a dummy world.

129. The author will probably continue by

(a) discussing the world of atomic and sub-atomic physics

(b) showing how most sciences are not science at all

(c) demonstrating the validity of the everyday world

(d) saying how little science has changed our experiance of the world

130. "hardy" (line 9) means

(a) difficult

(b) learned

(c) keen

(d) brave

131. "engender" (line 43) means

(a) establish

(b) create

(c) force

(d) encourage

132. "It" (line 50) refers to

(a) painted stage canvas

(b) hearsay knowledge

(c) The gnawing suspician

(d) The super-terrestrial

UNIVERSITI SAINS MALAYSIA
PUSAT BAHASA DAN TERJEMAHAN

ENGLISH LANGUAGE PLACEMENT TEST

FORM B

Part 4 - ½ hour

Part 4 is a test of your ability to summarize a passage of English.

Read quickly through the passage on the next page, then write a summary of the passage in English.

You should use between 80 - 150 words.

Use the separate answer sheet for your answer and for your rough work.

In these days of popular expositions, both written and broadcast, of Outlines, and of mammoth Guides to the Intelligent Man - guides through science, guides through economics, guides through philosophy, guides through chaos - the common reader cannot be unaware that the sciences in general and the physical sciences in particular have been developing rapidly and that in the course of this development certain changes, describable as 'revolutionary', have occured. These developments in science have a twofold interest. First, their results have given us information, often surprising, about the world we live in. Secondly, the following out of scientific method is in itself exciting, affording us the purest of all satisfactions - intellectual satisfaction. There is among common readers a genuine interest in scientific research, a desire to follow as far as a layman can what is being found and to understand the implications of these findings. Some of us are prepared to attempt to make the considerable intellectual effort required in order to understand even a non-technical exposition of recent developments in physics. The riting of such an exposition is undoubtedly difficult. It requires not only great powers of exposition but also an apprehension of the sort of difficulties the layman is likely to find and the skill to surmount them. We can hardly complain if there matters are not made entirely clear to us. Nevertheless, there are not a few scientists who have written books that to some extent satisfy our needs. Unfortunately, however, there are other famous scientists who do not seem to realize that their subject has an intrinsic interest for the common reader, and accordingly they seek to arouse his emotions, thereby inducing a frame of mind inimical to intellectual discernment. Popularizations of such a kind constitute a grave danger to thinking clearly. Possibly the authors themselves are at times wrought up to a pitch of emotional excitement, undaly impressed by the strangeness of their discoveries. I say 'unduly impressed' because, however strange may be the accounts of recent physical speculations, these physical speculations are themselves the development of the normal procedure of scientific method. The invention of new and more delicate scientific instruments has extended the physicist's range of experience; fresh mathematical techniques have had to be devised to deal with the discoveries thus made. It must not, however, be too hastily assumed that these new instruments and these new mathematical devices constitute in themselves a radical transformation of the nature of our knowledge. Some of our scientific guides, writing in moments of emotional exaltation, have found it easier to mystify the common reader than to enlighten him.

Appendix 3

Summary Statistics for Trial Versions of USM Placement Test

III. Classical Test Statistics: item analysis

**** ITEM ANALYSIS OF TEST A ****

SUMMARY LIST OF STATISTICS

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	POINT BISERIAL	(SUBSCALES..) 1	2	3	UNBIASED EQU. PT.BIS.	
A 1	0.62	0.50	0.43	0.39	0.37	0.33	0.41	1 C6
A 2	0.87	0.32	0.49	0.52	0.48	0.24	0.47	1 C36
A 3	0.87	0.28	0.40	0.38	0.33	0.31	0.38	1 D2
A 4	0.78	0.14	0.19	0.15	0.22	0.09	0.17	1 D43
A 5	0.40	0.23	0.20	0.18	0.15	0.12	0.17	1 D3
A 6	0.80	0.33	0.38	0.37	0.36	0.23	0.36	1 D49
A 7	0.90	0.26	0.41	0.38	0.40	0.26	0.39	1 D19
A 8	0.58	0.39	0.36	0.32	0.33	0.23	0.33	1 C45
A 9	0.86	0.31	0.49	0.51	0.44	0.28	0.47	1 D40
A 10	0.89	0.27	0.38	0.36	0.33	0.27	0.36	1 D15
A 11	0.90	0.27	0.41	0.45	0.40	0.18	0.40	1 C48
A 12	0.78	0.22	0.25	0.18	0.23	0.22	0.23	1 D29
A 13	0.56	0.72	0.59	0.57	0.50	0.46	0.58	1 D5
A 14	0.41	0.22	0.18	0.14	0.18	0.09	0.15	1 C9
A 15	0.12	0.22	0.27	0.25	0.18	0.26	0.26	1 C22
A 16	0.86	0.30	0.42	0.42	0.40	0.25	0.41	1 C29
A 17	0.48	0.57	0.49	0.41	0.40	0.44	0.47	1 C49
A 18	0.63	0.42	0.37	0.36	0.32	0.24	0.35	1 D41
A 19	0.83	0.42	0.53	0.53	0.51	0.31	0.51	1 C2
A 20	0.86	0.28	0.32	0.36	0.24	0.20	0.30	1 D16
A 21	0.97	0.08	0.23	0.21	0.24	0.12	0.22	1 C21
A 22	0.60	0.40	0.36	0.32	0.28	0.29	0.33	1 C13
A 23	0.74	0.34	0.35	0.34	0.30	0.24	0.33	1 D47
A 24	0.92	0.23	0.39	0.37	0.40	0.22	0.37	1 C23
A 25	0.41	0.37	0.33	0.29	0.26	0.27	0.31	1 D44
A 26	0.54	0.48	0.42	0.43	0.32	0.31	0.40	1 D12
A 27	0.57	0.58	0.48	0.47	0.42	0.31	0.46	1 D26
A 28	0.80	0.40	0.46	0.44	0.41	0.32	0.44	1 D39
A 29	0.89	0.26	0.41	0.41	0.43	0.22	0.40	1 C7
A 30	0.39	0.20	0.22	0.18	0.20	0.14	0.19	1 D50
A 31	0.67	0.28	0.25	0.24	0.17	0.18	0.23	1 D23
A 32	0.91	0.20	0.39	0.44	0.41	0.12	0.37	1 D31
A 33	0.80	0.42	0.49	0.45	0.45	0.35	0.47	1 C33
A 34	0.28	0.28	0.29	0.25	0.21	0.26	0.27	1 D6
A 35	0.87	0.29	0.43	0.40	0.39	0.31	0.41	1 D27
A 36	0.74	0.50	0.49	0.50	0.44	0.31	0.47	1 C50
A 37	0.46	0.02	0.05	0.01	-0.05	0.12	0.02	1 C30
A 38	0.31	0.08	0.10	0.09	0.01	0.10	0.07	1 C39
A 39	0.40	0.33	0.26	0.22	0.14	0.28	0.24	1 C46
A 40	0.62	0.58	0.51	0.47	0.43	0.40	0.49	1 C3

A	41	0.09	0.08	0.11	0.09	0.02	0.15	0.09	1 D38
A	42	0.70	0.51	0.48	0.43	0.43	0.37	0.46	1 C19
A	43	0.64	0.08	0.13	0.09	0.09	0.09	0.10	1 D46
A	44	0.89	0.26	0.38	0.45	0.34	0.16	0.37	1 D7
A	45	0.88	0.22	0.31	0.35	0.25	0.16	0.29	1 D9
A	46	0.40	0.52	0.46	0.41	0.39	0.39	0.44	1 C40
A	47	0.34	0.40	0.39	0.32	0.30	0.37	0.36	1 C31
A	48	0.53	0.39	0.32	0.30	0.25	0.24	0.30	1 C8
A	49	0.90	0.18	0.32	0.32	0.28	0.20	0.30	1 C24
A	50	0.10	0.09	0.17	0.12	0.19	0.08	0.15	1 C27
A	51	0.66	0.23	0.21	0.18	0.22	0.08	0.18	2 D73
A	52	0.87	0.36	0.49	0.49	0.49	0.28	0.48	2 D68
A	53	0.93	0.20	0.38	0.39	0.40	0.18	0.37	2 C52
A	54	0.78	0.47	0.54	0.52	0.48	0.39	0.52	2 C56
A	55	0.53	0.24	0.23	0.21	0.16	0.17	0.20	2 C71
A	56	0.92	0.21	0.38	0.35	0.41	0.21	0.36	2 D63
A	57	0.52	0.42	0.31	0.23	0.23	0.31	0.29	2 C55
A	58	0.91	0.24	0.42	0.40	0.42	0.25	0.41	2 C64
A	59	0.12	0.06	0.05	0.03	-0.02	0.07	0.03	2 D64
A	60	0.55	0.54	0.45	0.39	0.39	0.35	0.42	2 D69
A	61	0.68	0.60	0.53	0.54	0.42	0.38	0.51	2 C62
A	62	0.93	0.14	0.29	0.23	0.30	0.22	0.28	2 C66
A	63	0.45	0.49	0.40	0.33	0.26	0.41	0.37	2 C73
A	64	0.84	0.33	0.42	0.35	0.42	0.32	0.41	2 D59
A	65	0.23	0.00	0.09	0.08	0.05	0.04	0.07	2 D67
A	66	0.83	0.43	0.53	0.50	0.54	0.34	0.52	2 C54
A	67	0.83	0.22	0.21	0.20	0.15	0.17	0.19	2 D70
A	68	0.87	0.34	0.41	0.38	0.43	0.23	0.39	2 D65
A	69	0.55	0.48	0.41	0.36	0.35	0.33	0.39	2 C58
A	70	0.48	0.59	0.48	0.43	0.36	0.42	0.46	2 D60
A	71	0.77	0.14	0.17	0.17	0.13	0.09	0.15	2 D61
A	72	0.88	0.26	0.43	0.39	0.45	0.25	0.41	2 C68
A	73	0.62	0.64	0.54	0.52	0.45	0.41	0.52	2 D66
A	74	0.79	0.20	0.24	0.21	0.26	0.12	0.22	2 D74
A	75	0.72	0.30	0.33	0.25	0.36	0.23	0.31	2 C75
A	76	0.08	-0.04	-0.06	-0.10	-0.10	0.00	-0.08	2 D76
A	77	0.16	-0.07	-0.12	-0.14	-0.15	-0.07	-0.14	2 C77
A	78	0.82	0.23	0.31	0.24	0.32	0.22	0.29	2 C78-C81
A	79	0.72	0.44	0.47	0.38	0.49	0.34	0.45	2
A	80	0.72	0.43	0.44	0.35	0.43	0.35	0.42	2
A	81	0.72	0.40	0.41	0.32	0.43	0.30	0.39	2
A	82	0.85	0.36	0.45	0.48	0.43	0.24	0.44	2 D82-D84
A	83	0.70	0.40	0.41	0.39	0.39	0.26	0.39	2
A	84	0.81	0.13	0.17	0.13	0.13	0.12	0.14	2
A	85	0.88	0.13	0.23	0.19	0.24	0.14	0.22	2 C85-C94
A	86	0.86	0.12	0.19	0.16	0.22	0.07	0.17	2
A	87	0.83	0.07	0.10	0.06	0.11	0.03	0.07	2
A	88	0.92	0.11	0.23	0.21	0.29	0.08	0.22	2
A	89	0.96	0.04	0.19	0.14	0.26	0.09	0.18	2
A	90	0.96	0.07	0.25	0.19	0.33	0.11	0.24	2
A	91	0.96	0.08	0.20	0.19	0.28	0.04	0.19	2
A	92	0.94	0.09	0.24	0.18	0.31	0.12	0.22	2
A	93	0.96	0.10	0.30	0.26	0.33	0.20	0.29	2
A	94	0.98	0.06	0.30	0.26	0.34	0.19	0.29	2
A	95	0.72	0.42	0.44	0.41	0.40	0.29	0.42	2 D95-D100
A	96	0.93	0.19	0.39	0.36	0.42	0.24	0.38	2
A	97	0.50	0.41	0.36	0.26	0.34	0.30	0.33	2

A 98	0.96	0.13	0.32	0.29	0.31	0.23	0.31	2	
A 99	0.33	-0.07	-0.06	-0.09	-0.10	-0.04	-0.09	2	
A100	0.84	0.34	0.41	0.32	0.43	0.32	0.40	2	
A101	0.67	0.48	0.38	0.33	0.32	0.32	0.36	3	C109-C116
A102	0.36	0.22	0.27	0.21	0.18	0.27	0.25	3	
A103	0.35	0.13	0.13	0.07	0.09	0.11	0.10	3	
A104	0.80	0.19	0.25	0.20	0.22	0.18	0.22	3	
A105	0.65	0.32	0.32	0.22	0.33	0.23	0.29	3	
A106	0.67	0.54	0.48	0.37	0.42	0.45	0.46	3	
A107	0.36	0.28	0.26	0.21	0.16	0.26	0.23	3	
A108	0.58	0.46	0.40	0.35	0.32	0.33	0.37	3	
A109	0.14	0.09	0.14	0.09	0.10	0.16	0.13	3	C117-C124
A110	0.56	0.11	0.06	-0.04	0.06	0.08	0.03	3	
A111	0.41	0.04	0.03	-0.01	-0.05	0.08	0.00	3	
A112	0.39	0.13	0.16	0.11	0.12	0.12	0.13	3	
A113	0.71	0.33	0.33	0.26	0.30	0.28	0.31	3	
A114	0.38	0.49	0.42	0.36	0.28	0.45	0.40	3	
A115	0.65	0.44	0.38	0.28	0.32	0.39	0.36	3	
A116	0.36	0.44	0.39	0.34	0.24	0.42	0.37	3	
A117	0.55	0.21	0.18	0.08	0.17	0.17	0.15	3	C125-C134
A118	0.86	0.26	0.35	0.29	0.29	0.32	0.34	3	
A119	0.68	0.60	0.54	0.45	0.50	0.45	0.53	3	
A120	0.78	0.33	0.39	0.31	0.33	0.36	0.37	3	
A121	0.33	0.52	0.46	0.37	0.38	0.44	0.44	3	
A122	0.86	0.30	0.40	0.32	0.40	0.30	0.38	3	
A123	0.54	0.11	0.14	0.07	0.13	0.11	0.11	3	
A124	0.39	0.37	0.35	0.28	0.26	0.36	0.33	3	
A125	0.38	0.34	0.34	0.26	0.26	0.33	0.31	3	
A126	0.59	0.39	0.33	0.27	0.25	0.29	0.30	3	
A127	0.42	0.48	0.41	0.35	0.31	0.37	0.38	3	
A128	0.58	0.38	0.32	0.26	0.22	0.32	0.29	3	
A129	0.20	0.12	0.14	0.11	0.06	0.14	0.12	3	
A130	0.51	0.41	0.35	0.28	0.25	0.34	0.32	3	
A131	0.30	0.11	0.10	0.05	0.01	0.16	0.08	3	
A132	0.38	0.22	0.20	0.20	0.10	0.18	0.18	3	
A133	0.26	0.08	0.07	0.06	-0.01	0.07	0.04	3	
A134	0.30	0.02	0.03	0.03	-0.09	0.08	0.01	3	
A135	0.21	0.04	0.05	0.03	-0.02	0.05	0.03	3	
A136	0.55	0.47	0.38	0.33	0.25	0.36	0.35	3	
A137	0.16	0.12	0.19	0.12	0.15	0.21	0.17	3	
A138	0.54	0.49	0.41	0.34	0.35	0.35	0.39	3	
A139	0.34	0.37	0.37	0.29	0.30	0.35	0.35	3	
A140	0.25	0.11	0.10	0.08	0.02	0.10	0.08	3	

LISTED DISCRIMINATION INDICES ARE PROP(GROUP 3) - PROP(GROUP 1)
LISTED CORRELATIONS ARE POINT-BISERIALS BETWEEN ITEM SCORE
(0 OR 1) AND TOTAL SCORE.
SUB-SCALE CORRELATIONS ARE (UNBIASED) POINT-BISERIALS
BETWEEN ITEM SCORE AND SUB-SCALE SCORE.

TABLE OF INTERCORRELATIONS

2	0.792	1.000

3	0.657	0.629	1.000
	1	2	3

NUMBER OF CANDIDATES = 269

TOTAL SCORE MEAN =	87.70	STANDARD DEVIATION=	18.07
SUB 1 SCORE MEAN =	32.37	STANDARD DEVIATION =	7.66
SUB 2 SCORE MEAN =	36.33	STANDARD DEVIATION =	6.49
SUB 3 SCORE MEAN =	18.99	STANDARD DEVIATION =	6.06

INTERNAL CONSISTENCY	RELIABILITY (KR20)	=	0.932
SUBTEST 1	"	" =	0.871
SUBTEST 2	"	" =	0.845
SUBTEST 3	"	" =	0.788

**** ITEM ANALYSIS OF TEST B ****

SUMMARY LIST OF STATISTICS

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	POINT BISERIAL	(SUBSCALES..) 1	2	3	UNBIASED EQU. PT.BIS.	
B 1	0.59	0.70	0.58	0.53	0.52	0.49	0.56	1 C34
B 2	0.28	0.37	0.38	0.31	0.33	0.37	0.36	1 D35
B 3	0.48	0.54	0.46	0.44	0.36	0.39	0.44	1 C14
B 4	0.71	0.24	0.28	0.25	0.25	0.20	0.26	1 D42
B 5	0.73	0.34	0.31	0.31	0.26	0.22	0.30	1 D28
B 6	0.35	0.59	0.51	0.44	0.46	0.49	0.50	1 D21
B 7	0.81	0.45	0.45	0.47	0.38	0.31	0.43	1 D17
B 8	0.42	0.37	0.31	0.30	0.29	0.16	0.29	1 C44
B 9	0.64	0.55	0.48	0.49	0.36	0.42	0.47	1 C4
B 10	0.67	0.68	0.60	0.61	0.53	0.43	0.59	1 D4
B 11	0.81	0.43	0.47	0.47	0.41	0.35	0.46	1 D18
B 12	0.75	0.39	0.38	0.38	0.32	0.26	0.36	1 C17
B 13	0.54	0.60	0.52	0.50	0.45	0.39	0.50	1 D32
B 14	0.51	0.02	0.05	0.02	0.02	0.02	0.02	1 C38
B 15	0.11	0.10	0.14	0.11	0.14	0.09	0.13	1 C25
B 16	0.78	0.17	0.19	0.20	0.18	0.07	0.18	1 C43
B 17	0.72	0.50	0.48	0.47	0.45	0.32	0.46	1 C42
B 18	0.68	0.55	0.51	0.47	0.46	0.40	0.49	1 C32
B 19	0.84	0.34	0.43	0.46	0.34	0.32	0.42	1 D20
B 20	0.75	0.57	0.58	0.61	0.52	0.40	0.57	1 C5
B 21	0.45	0.39	0.37	0.32	0.31	0.32	0.35	1 C20
B 22	0.75	0.59	0.59	0.61	0.54	0.39	0.58	1 C35
B 23	0.51	0.66	0.55	0.52	0.47	0.46	0.53	1 C15
B 24	0.78	0.12	0.16	0.12	0.11	0.18	0.14	1 D10
B 25	0.41	0.56	0.52	0.45	0.43	0.51	0.50	1 C10
B 26	0.83	0.37	0.43	0.44	0.37	0.29	0.41	1 D8
B 27	0.47	0.44	0.39	0.39	0.31	0.28	0.37	1 C1
B 28	0.71	0.59	0.55	0.54	0.52	0.38	0.53	1 C16
B 29	0.55	0.70	0.58	0.56	0.52	0.46	0.57	1 D36
B 30	0.73	0.46	0.48	0.49	0.41	0.35	0.47	1 D34
B 31	0.77	0.59	0.59	0.60	0.54	0.40	0.58	1 C28
B 32	0.25	0.12	0.16	0.14	0.13	0.11	0.14	1 D14
B 33	0.67	0.34	0.33	0.29	0.25	0.32	0.31	1 C26
B 34	0.45	0.68	0.59	0.56	0.49	0.52	0.57	1 C12
B 35	0.76	0.43	0.43	0.45	0.38	0.27	0.42	1 D13
B 36	0.84	0.22	0.30	0.33	0.24	0.18	0.28	1 D1
B 37	0.57	0.56	0.50	0.50	0.41	0.39	0.48	1 D30
B 38	0.93	0.18	0.35	0.35	0.33	0.22	0.34	1 D22
B 39	0.56	0.54	0.48	0.41	0.44	0.41	0.46	1 C41
B 40	0.89	0.29	0.39	0.39	0.32	0.31	0.38	1 C18
B 41	0.82	0.44	0.49	0.48	0.45	0.36	0.48	1 D33
B 42	0.89	0.27	0.39	0.42	0.33	0.25	0.38	1 C11
B 43	0.59	0.74	0.63	0.62	0.59	0.45	0.62	1 D45
B 44	0.96	0.10	0.27	0.29	0.22	0.17	0.26	1 C37
B 45	0.82	0.28	0.30	0.30	0.21	0.24	0.28	1 D25

B 46	0.58	0.59	0.49	0.49	0.44	0.34	0.48	1	D24
B 47	0.33	0.18	0.22	0.14	0.14	0.32	0.20	1	D11
B 48	0.74	0.51	0.52	0.55	0.40	0.42	0.51	1	C47
B 49	0.77	0.38	0.36	0.34	0.32	0.27	0.34	1	D37
B 50	0.84	0.45	0.50	0.53	0.44	0.30	0.49	1	D48
B 51	0.35	0.54	0.45	0.40	0.42	0.37	0.44	2	D56
B 52	0.58	0.34	0.27	0.27	0.24	0.16	0.25	2	C72
B 53	0.41	0.35	0.34	0.31	0.31	0.24	0.31	2	C51
B 54	0.55	0.44	0.39	0.33	0.33	0.37	0.37	2	D58
B 55	0.59	0.49	0.42	0.41	0.36	0.31	0.40	2	C57
B 56	0.39	0.18	0.17	0.14	0.09	0.17	0.14	2	D53
B 57	0.55	0.44	0.34	0.29	0.33	0.28	0.32	2	D71
B 58	0.70	0.54	0.49	0.47	0.41	0.38	0.47	2	C53
B 59	0.60	0.06	0.07	0.05	0.02	0.06	0.05	2	D54
B 60	0.04	-0.06	-0.16	-0.18	-0.17	-0.09	-0.16	2	C69
B 61	0.42	0.32	0.30	0.27	0.25	0.24	0.28	2	C59
B 62	0.69	0.60	0.54	0.54	0.47	0.39	0.52	2	C61
B 63	0.12	-0.01	0.03	0.01	0.02	0.02	0.02	2	D55
B 64	0.68	0.55	0.51	0.48	0.43	0.43	0.49	2	C60
B 65	0.93	0.17	0.29	0.30	0.24	0.19	0.28	2	C70
B 66	0.76	0.54	0.54	0.54	0.48	0.37	0.52	2	D72
B 67	0.56	0.24	0.24	0.21	0.18	0.22	0.22	2	D57
B 68	0.65	0.45	0.40	0.39	0.40	0.24	0.39	2	C65
B 69	0.80	0.49	0.51	0.50	0.47	0.35	0.50	2	C67
B 70	0.76	0.52	0.53	0.54	0.47	0.37	0.52	2	C63
B 71	0.26	0.39	0.42	0.32	0.31	0.53	0.41	2	D62
B 72	0.85	0.24	0.30	0.30	0.24	0.21	0.28	2	D51
B 73	0.74	0.51	0.49	0.48	0.43	0.36	0.47	2	D52
B 74	0.62	0.34	0.34	0.30	0.30	0.28	0.32	2	C74
B 75	0.75	0.43	0.39	0.38	0.36	0.26	0.38	2	D75
B 76	0.61	0.21	0.16	0.12	0.14	0.13	0.14	2	C76
B 77	0.34	0.05	0.10	0.06	0.08	0.06	0.08	2	D77
B 78	0.95	0.13	0.29	0.29	0.25	0.20	0.28	2	C82-C84
B 79	0.76	0.35	0.35	0.29	0.36	0.28	0.34	2	
B 80	0.75	0.39	0.39	0.34	0.39	0.30	0.38	2	
B 81	0.67	0.44	0.41	0.34	0.42	0.31	0.39	2	D78-D81
B 82	0.71	0.35	0.34	0.27	0.36	0.25	0.32	2	
B 83	0.64	0.49	0.45	0.39	0.45	0.34	0.43	2	
B 84	0.78	0.30	0.35	0.30	0.36	0.26	0.33	2	
B 85	0.45	-0.23	-0.26	-0.24	-0.27	-0.26	-0.28	2	D85-D94
B 86	0.64	0.34	0.32	0.29	0.25	0.29	0.30	2	
B 87	0.61	0.10	0.09	0.04	0.11	0.03	0.07	2	
B 88	0.60	0.12	0.11	0.05	0.15	0.06	0.09	2	
B 89	0.38	0.34	0.30	0.21	0.37	0.17	0.27	2	
B 90	0.62	0.50	0.42	0.33	0.46	0.32	0.40	2	
B 91	0.69	0.50	0.46	0.39	0.49	0.31	0.44	2	
B 92	0.69	0.45	0.43	0.34	0.47	0.30	0.41	2	
B 93	0.27	0.29	0.31	0.23	0.34	0.24	0.29	2	
B 94	0.31	0.16	0.19	0.12	0.20	0.15	0.17	2	
B 95	0.51	0.30	0.26	0.23	0.23	0.19	0.24	2	D95-D10
B 96	0.47	0.65	0.55	0.50	0.48	0.47	0.53	2	
B 97	0.56	0.12	0.13	0.10	0.09	0.12	0.11	2	
B 98	0.65	0.35	0.32	0.27	0.29	0.25	0.30	2	
B 99	0.71	0.55	0.49	0.44	0.45	0.42	0.47	2	
B100	0.86	0.24	0.33	0.31	0.30	0.24	0.31	2	
B101	0.36	0.68	0.62	0.55	0.53	0.59	0.60	3	D117-D
B102	0.55	0.33	0.30	0.27	0.25	0.26	0.28	3	

B103	0.81	0.30	0.28	0.25	0.24	0.23	0.26	3
B104	0.74	0.49	0.45	0.42	0.39	0.39	0.44	3
B105	0.49	0.59	0.50	0.44	0.46	0.42	0.48	3
B106	0.57	0.33	0.34	0.31	0.28	0.30	0.32	3
B107	0.41	0.39	0.37	0.32	0.29	0.36	0.35	3
B108	0.32	0.12	0.12	0.08	0.05	0.16	0.09	3
B109	0.58	0.10	0.15	0.11	0.09	0.18	0.13	3 D101-D1
B110	0.37	0.28	0.23	0.17	0.21	0.22	0.21	3
B111	0.31	-0.11	-0.08	-0.07	-0.07	-0.16	-0.10	3
B112	0.36	0.32	0.31	0.24	0.23	0.34	0.29	3
B113	0.36	0.18	0.19	0.17	0.10	0.20	0.17	3
B114	0.65	0.13	0.17	0.12	0.10	0.21	0.15	3
B115	0.44	0.23	0.24	0.19	0.19	0.22	0.21	3
B116	0.51	0.49	0.40	0.33	0.36	0.38	0.38	3
B117	0.29	0.38	0.35	0.27	0.31	0.35	0.33	3
B118	0.39	0.56	0.52	0.48	0.42	0.46	0.50	3
B119	0.92	0.12	0.23	0.23	0.17	0.18	0.22	3
B120	0.56	0.55	0.50	0.44	0.47	0.41	0.48	3
B121	0.22	-0.16	-0.17	-0.19	-0.14	-0.19	-0.19	3
B122	0.49	0.70	0.57	0.51	0.51	0.51	0.56	3
B123	0.57	0.61	0.53	0.53	0.47	0.38	0.51	3
B124	0.44	0.10	0.12	0.10	0.06	0.09	0.09	3
B125	0.57	-0.06	-0.03	-0.03	-0.04	-0.08	-0.05	3
B126	0.39	0.09	0.10	0.04	0.05	0.16	0.08	3
B127	0.31	0.10	0.05	0.01	0.03	0.05	0.03	3
B128	0.22	0.11	0.17	0.06	0.15	0.26	0.15	3
B129	0.25	0.10	0.07	0.04	0.03	0.07	0.05	3
B130	0.49	0.48	0.43	0.40	0.34	0.36	0.41	3
B131	0.36	0.23	0.19	0.17	0.19	0.08	0.17	3
B132	0.51	0.44	0.43	0.41	0.37	0.35	0.42	3
B133	0.33	0.32	0.31	0.25	0.28	0.27	0.29	3 D133-D1
B134	0.15	0.09	0.16	0.14	0.12	0.15	0.14	3
B135	0.36	0.60	0.52	0.46	0.47	0.47	0.51	3
B136	0.22	0.21	0.20	0.14	0.18	0.20	0.18	3
B137	0.36	0.24	0.27	0.19	0.27	0.22	0.25	3
B138	0.27	0.44	0.47	0.40	0.37	0.49	0.45	3
B139	0.17	-0.06	-0.05	-0.04	-0.08	-0.06	-0.06	3
B140	0.24	0.01	0.02	0.01	-0.02	-0.01	-0.00	3

LISTED DISCRIMINATION INDICES ARE PROP(GROUP 3) - PROP(GROUP 1)
LISTED CORRELATIONS ARE POINT-BISERIALS BETWEEN ITEM SCORE
(0 OR 1) AND TOTAL SCORE.
SUB-SCALE CORRELATIONS ARE (UNBIASED) POINT-BISERIALS
BETWEEN ITEM SCORE AND SUB-SCALE SCORE.

TABLE OF INTERCORRELATIONS

2	0.817	1.000	
3	0.728	0.708	1.000
	1	2	3

NUMBER OF CANDIDATES = 245

TOTAL SCORE MEAN =	78.82	STANDARD DEVIATION =	21.47
SUB 1 SCORE MEAN =	32.37	STANDARD DEVIATION =	9.76
SUB 2 SCORE MEAN =	29.57	STANDARD DEVIATION =	7.79
SUB 3 SCORE MEAN =	16.89	STANDARD DEVIATION =	5.82

INTERNAL CONSISTENCY RELIABILITY (KR20) =		0.945
SUBTEST 1 " " =		0.917
SUBTEST 2 " " =		0.848
SUBTEST 3 " " =		0.765

**** ITEM ANALYSIS OF TEST C ****

SUMMARY LIST OF STATISTICS

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	POINT BISERIAL	(SUBSCALES..) 1	2	3	UNBIASED PT.BIS.	
C 1	0.55	0.33	0.30	0.25	0.24	0.29	0.28	1
C 2	0.79	0.40	0.43	0.40	0.39	0.36	0.42	1
C 3	0.51	0.51	0.43	0.40	0.35	0.41	0.42	1
C 4	0.69	0.60	0.52	0.49	0.46	0.44	0.50	1
C 5	0.74	0.58	0.56	0.52	0.54	0.45	0.54	1
C 6	0.57	0.64	0.56	0.58	0.46	0.48	0.55	1
C 7	0.82	0.40	0.49	0.45	0.48	0.39	0.48	1
C 8	0.37	0.37	0.33	0.31	0.27	0.29	0.31	1
C 9	0.41	0.22	0.20	0.21	0.16	0.13	0.18	1
C 10	0.43	0.66	0.54	0.52	0.46	0.48	0.53	1
C 11	0.88	0.25	0.35	0.34	0.34	0.25	0.34	1
C 12	0.41	0.61	0.52	0.48	0.41	0.50	0.50	1
C 13	0.52	0.45	0.40	0.35	0.37	0.35	0.38	1
C 14	0.42	0.46	0.43	0.39	0.40	0.36	0.41	1
C 15	0.41	0.58	0.53	0.51	0.42	0.50	0.51	1
C 16	0.62	0.63	0.54	0.51	0.46	0.50	0.53	1
C 17	0.76	0.43	0.44	0.42	0.40	0.36	0.43	1
C 18	0.88	0.28	0.39	0.37	0.37	0.29	0.37	1
C 19	0.57	0.63	0.53	0.52	0.46	0.45	0.52	1
C 20	0.26	0.35	0.37	0.32	0.31	0.35	0.35	1
C 21	0.88	0.22	0.34	0.30	0.35	0.25	0.32	1
C 22	0.31	0.53	0.48	0.45	0.40	0.46	0.47	1
C 23	0.84	0.32	0.39	0.36	0.37	0.33	0.38	1
C 24	0.87	0.21	0.30	0.31	0.23	0.23	0.28	1
C 25	0.36	0.30	0.30	0.28	0.26	0.25	0.28	1
C 26	0.75	0.25	0.25	0.24	0.17	0.26	0.24	1
C 27	0.12	0.05	0.06	0.06	-0.02	0.09	0.04	1
C 28	0.74	0.54	0.52	0.48	0.46	0.47	0.50	1
C 29	0.78	0.40	0.44	0.44	0.39	0.33	0.42	1
C 30	0.48	-0.08	-0.05	-0.06	-0.08	-0.06	-0.07	1
C 31	0.25	0.24	0.29	0.26	0.22	0.28	0.27	1
C 32	0.68	0.53	0.48	0.46	0.43	0.38	0.46	1
C 33	0.56	0.70	0.61	0.58	0.53	0.55	0.60	1
C 34	0.49	0.76	0.64	0.61	0.57	0.55	0.62	1
C 35	0.79	0.50	0.52	0.54	0.48	0.37	0.50	1
C 36	0.78	0.52	0.55	0.54	0.52	0.40	0.53	1
C 37	0.93	0.18	0.34	0.30	0.35	0.26	0.33	1
C 38	0.63	0.29	0.26	0.21	0.25	0.23	0.24	1
C 39	0.31	0.08	0.13	0.12	0.11	0.07	0.11	1
C 40	0.36	0.47	0.45	0.41	0.35	0.44	0.43	1
C 41	0.50	0.46	0.38	0.33	0.30	0.37	0.36	1
C 42	0.71	0.43	0.42	0.40	0.37	0.36	0.41	1
C 43	0.75	0.26	0.28	0.27	0.25	0.21	0.26	1
C 44	0.37	0.32	0.31	0.28	0.27	0.27	0.29	1
C 45	0.52	0.36	0.36	0.35	0.29	0.29	0.34	1

C 46	0.35	0.13	0.15	0.12	0.09	0.16	0.13	1
C 47	0.75	0.42	0.43	0.39	0.42	0.35	0.42	1
C 48	0.83	0.39	0.43	0.41	0.41	0.34	0.42	1
C 49	0.41	0.42	0.37	0.33	0.32	0.33	0.35	1
C 50	0.64	0.57	0.52	0.51	0.43	0.46	0.51	1
C 51	0.39	0.41	0.38	0.38	0.31	0.33	0.37	2
C 52	0.87	0.35	0.48	0.46	0.50	0.35	0.47	2
C 53	0.59	0.45	0.43	0.39	0.36	0.38	0.41	2
C 54	0.78	0.45	0.48	0.43	0.47	0.38	0.46	2
C 55	0.51	0.17	0.22	0.21	0.18	0.14	0.19	2
C 56	0.66	0.71	0.62	0.60	0.53	0.57	0.61	2
C 57	0.36	0.60	0.50	0.47	0.40	0.49	0.49	2
C 58	0.58	0.41	0.36	0.33	0.31	0.32	0.35	2
C 59	0.45	0.23	0.24	0.21	0.19	0.23	0.22	2
C 60	0.66	0.51	0.45	0.43	0.42	0.35	0.43	2
C 61	0.64	0.52	0.52	0.51	0.49	0.38	0.50	2
C 62	0.55	0.72	0.58	0.57	0.48	0.53	0.57	2
C 63	0.68	0.50	0.52	0.50	0.48	0.41	0.50	2
C 64	0.87	0.38	0.50	0.46	0.50	0.41	0.49	2
C 65	0.64	0.53	0.51	0.48	0.48	0.40	0.49	2
C 66	0.86	0.28	0.39	0.35	0.41	0.30	0.38	2
C 67	0.78	0.53	0.58	0.55	0.56	0.44	0.56	2
C 68	0.84	0.38	0.43	0.39	0.41	0.35	0.41	2
C 69	0.03	-0.09	-0.21	-0.22	-0.20	-0.19	-0.22	2
C 70	0.90	0.23	0.37	0.31	0.44	0.24	0.36	2
C 71	0.74	0.35	0.33	0.26	0.34	0.29	0.31	2
C 72	0.57	0.50	0.41	0.37	0.37	0.37	0.39	2
C 73	0.39	0.28	0.25	0.25	0.15	0.24	0.23	2
C 74	0.38	0.11	0.12	0.08	0.12	0.10	0.10	2
C 75	0.56	0.24	0.23	0.20	0.22	0.16	0.21	2
C 76	0.54	0.12	0.18	0.14	0.16	0.15	0.16	2
C 77	0.14	-0.07	-0.07	-0.04	-0.09	-0.10	-0.08	2
C 78	0.74	0.29	0.31	0.25	0.33	0.24	0.29	2
C 79	0.64	0.36	0.34	0.27	0.38	0.25	0.32	2
C 80	0.63	0.43	0.41	0.37	0.42	0.30	0.39	2
C 81	0.61	0.41	0.39	0.34	0.40	0.30	0.37	2
C 82	0.88	0.16	0.26	0.24	0.29	0.14	0.24	2
C 83	0.60	0.36	0.31	0.28	0.32	0.20	0.29	2
C 84	0.57	0.41	0.34	0.31	0.36	0.23	0.32	2
C 85	0.87	0.11	0.18	0.15	0.15	0.15	0.16	2
C 86	0.80	0.22	0.26	0.21	0.32	0.15	0.25	2
C 87	0.80	0.11	0.13	0.09	0.16	0.06	0.11	2
C 88	0.92	0.13	0.26	0.22	0.28	0.20	0.25	2
C 89	0.96	0.11	0.31	0.22	0.41	0.21	0.30	2
C 90	0.94	0.14	0.33	0.24	0.44	0.22	0.32	2
C 91	0.93	0.12	0.26	0.19	0.30	0.21	0.25	2
C 92	0.95	0.09	0.24	0.21	0.27	0.16	0.23	2
C 93	0.95	0.13	0.36	0.33	0.39	0.24	0.35	2
C 94	0.97	0.09	0.30	0.22	0.38	0.22	0.29	2
C 95	0.51	0.41	0.35	0.32	0.29	0.32	0.33	2
C 96	0.46	0.63	0.53	0.50	0.44	0.48	0.51	2
C 97	0.47	0.00	0.04	0.05	-0.01	0.01	0.02	2
C 98	0.76	0.25	0.27	0.22	0.30	0.19	0.25	2
C 99	0.57	0.46	0.42	0.39	0.37	0.36	0.40	2
C100	0.83	0.26	0.29	0.29	0.25	0.22	0.27	2
C101	0.73	0.46	0.46	0.43	0.38	0.43	0.44	3
C102	0.85	0.36	0.46	0.42	0.45	0.36	0.45	3

C103	0.46	0.74	0.62	0.56	0.52	0.62	0.60	3
C104	0.89	0.29	0.44	0.38	0.45	0.34	0.42	3
C105	0.42	0.52	0.42	0.40	0.33	0.39	0.40	3
C106	0.40	0.58	0.49	0.48	0.37	0.48	0.47	3
C107	0.46	0.64	0.52	0.50	0.43	0.48	0.51	3
C108	0.36	0.30	0.29	0.28	0.24	0.23	0.27	3
C109	0.64	0.46	0.38	0.35	0.29	0.36	0.36	3
C110	0.33	0.26	0.22	0.22	0.19	0.14	0.20	3
C111	0.29	0.08	0.06	0.03	0.01	0.07	0.04	3
C112	0.78	0.35	0.32	0.31	0.25	0.29	0.31	3
C113	0.44	0.47	0.39	0.37	0.31	0.36	0.37	3
C114	0.54	0.61	0.51	0.47	0.43	0.46	0.49	3
C115	0.30	0.14	0.18	0.17	0.08	0.20	0.16	3
C116	0.42	0.37	0.36	0.31	0.30	0.34	0.34	3
C117	0.13	0.05	0.06	0.05	0.01	0.06	0.04	3
C118	0.45	0.04	0.03	-0.01	-0.04	0.08	0.01	3
C119	0.37	-0.04	-0.01	-0.03	-0.06	-0.00	-0.03	3
C120	0.38	0.15	0.16	0.14	0.14	0.12	0.14	3
C121	0.66	0.36	0.34	0.32	0.29	0.29	0.32	3
C122	0.29	0.27	0.32	0.33	0.24	0.26	0.30	3
C123	0.64	0.26	0.23	0.20	0.19	0.18	0.21	3
C124	0.32	0.42	0.43	0.44	0.34	0.37	0.42	3
C125	0.50	0.40	0.36	0.34	0.31	0.30	0.34	3
C126	0.75	0.35	0.32	0.29	0.24	0.33	0.31	3
C127	0.57	0.70	0.58	0.54	0.51	0.53	0.56	3
C128	0.64	0.54	0.46	0.41	0.40	0.45	0.45	3
C129	0.27	0.53	0.51	0.45	0.41	0.54	0.50	3
C130	0.76	0.40	0.42	0.36	0.38	0.39	0.40	3
C131	0.43	0.23	0.18	0.14	0.17	0.14	0.16	3
C132	0.29	0.24	0.21	0.20	0.09	0.27	0.20	3
C133	0.68	0.28	0.27	0.19	0.29	0.22	0.25	3
C134	0.36	0.34	0.31	0.29	0.25	0.27	0.29	3
C135	0.17	0.11	0.12	0.09	0.04	0.16	0.10	3
C136	0.35	0.30	0.28	0.28	0.19	0.24	0.26	3
C137	0.20	0.15	0.17	0.14	0.14	0.14	0.15	3
C138	0.55	0.45	0.39	0.37	0.33	0.34	0.37	3
C139	0.57	0.43	0.34	0.26	0.31	0.34	0.32	3
C140	0.67	0.41	0.40	0.35	0.33	0.38	0.38	3

LISTED DISCRIMINATION INDICES ARE PROP(GROUP 3) - PROP(GROUP 1)
LISTED CORRELATIONS ARE POINT-BISERIALS BETWEEN ITEM SCORE
(0 OR 1) AND TOTAL SCORE.
SUB-SCALE CORRELATIONS ARE (UNBIASED) POINT-BISERIALS
BETWEEN ITEM SCORE AND SUB-SCALE SCORE.

TABLE OF INTERCORRELATIONS

2	0.831	1.000	
3	0.825	0.732	1.000
	1	2	3

NUMBER OF CANDIDATES = 276

TOTAL SCORE MEAN =	81.92	STANDARD DEVIATION =	21.99
SUB 1 SCORE MEAN =	29.33	STANDARD DEVIATION =	9.25
SUB 2 SCORE MEAN =	33.32	STANDARD DEVIATION =	7.69
SUB 3 SCORE MEAN =	19.28	STANDARD DEVIATION =	6.65

INTERNAL CONSISTENCY RELIABILITY (KR20) =		0.950
SUBTEST 1	" " =	0.900
SUBTEST 2	" " =	0.867
SUBTEST 3	" " =	0.828

**** ITEM ANALYSIS OF TEST D ****

SUMMARY LIST OF STATISTICS

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	POINT BISERIAL	(SUBSCALES..) 1	2	3	UNBIASED PT.BIS.	
D 1	0.86	0.22	0.30	0.29	0.24	0.23	0.28	1
D 2	0.79	0.39	0.44	0.41	0.42	0.31	0.43	1
D 3	0.27	0.31	0.32	0.28	0.24	0.30	0.30	1
D 4	0.62	0.58	0.49	0.46	0.41	0.41	0.47	1
D 5	0.54	0.53	0.46	0.43	0.41	0.37	0.45	1
D 6	0.20	0.15	0.19	0.11	0.16	0.20	0.17	1
D 7	0.81	0.48	0.53	0.52	0.49	0.37	0.51	1
D 8	0.79	0.45	0.46	0.44	0.40	0.34	0.44	1
D 9	0.90	0.22	0.32	0.33	0.24	0.25	0.31	1
D 10	0.74	0.02	0.09	0.08	0.03	0.06	0.06	1
D 11	0.29	0.07	0.16	0.11	0.06	0.24	0.14	1
D 12	0.50	0.37	0.35	0.31	0.28	0.31	0.33	1
D 13	0.74	0.35	0.36	0.34	0.29	0.27	0.34	1
D 14	0.26	0.03	0.05	0.07	0.01	-0.01	0.03	1
D 15	0.80	0.45	0.45	0.46	0.39	0.31	0.43	1
D 16	0.80	0.38	0.42	0.42	0.39	0.26	0.40	1
D 17	0.73	0.52	0.51	0.52	0.45	0.36	0.50	1
D 18	0.83	0.34	0.36	0.35	0.34	0.24	0.35	1
D 19	0.89	0.25	0.34	0.30	0.34	0.25	0.33	1
D 20	0.79	0.24	0.23	0.22	0.20	0.15	0.21	1
D 21	0.38	0.49	0.48	0.43	0.40	0.43	0.46	1
D 22	0.91	0.21	0.33	0.32	0.29	0.23	0.31	1
D 23	0.63	0.18	0.17	0.14	0.13	0.12	0.14	1
D 24	0.60	0.60	0.49	0.51	0.41	0.35	0.48	1
D 25	0.79	0.20	0.20	0.19	0.13	0.18	0.18	1
D 26	0.55	0.51	0.44	0.39	0.39	0.35	0.42	1
D 27	0.67	0.57	0.54	0.56	0.43	0.40	0.52	1
D 28	0.77	0.45	0.40	0.36	0.35	0.34	0.38	1
D 29	0.70	0.37	0.37	0.34	0.31	0.31	0.35	1
D 30	0.56	0.48	0.43	0.43	0.34	0.34	0.41	1
D 31	0.76	0.54	0.52	0.52	0.47	0.37	0.51	1
D 32	0.52	0.66	0.57	0.56	0.46	0.46	0.55	1
D 33	0.84	0.43	0.50	0.51	0.48	0.30	0.49	1
D 34	0.83	0.33	0.40	0.37	0.37	0.29	0.38	1
D 35	0.40	0.37	0.36	0.32	0.31	0.28	0.34	1
D 36	0.45	0.79	0.66	0.65	0.56	0.51	0.64	1
D 37	0.65	0.35	0.32	0.31	0.30	0.17	0.29	1
D 38	0.10	0.02	0.09	0.07	0.05	0.09	0.07	1
D 39	0.65	0.52	0.46	0.46	0.39	0.33	0.44	1
D 40	0.73	0.55	0.53	0.47	0.50	0.40	0.51	1
D 41	0.50	0.43	0.42	0.46	0.30	0.31	0.40	1
D 42	0.64	0.27	0.27	0.23	0.17	0.27	0.24	1
D 43	0.71	0.34	0.37	0.41	0.32	0.17	0.35	1
D 44	0.25	0.35	0.32	0.33	0.26	0.22	0.30	1
D 45	0.54	0.71	0.60	0.57	0.52	0.48	0.58	1

D 46	0.50	0.15	0.16	0.14	0.11	0.12	0.14	1
D 47	0.69	0.42	0.40	0.41	0.32	0.29	0.38	1
D 48	0.84	0.38	0.48	0.50	0.46	0.28	0.47	1
D 49	0.68	0.38	0.34	0.33	0.26	0.28	0.32	1
D 50	0.33	0.21	0.27	0.26	0.19	0.23	0.25	1
D 51	0.83	0.31	0.40	0.36	0.40	0.26	0.38	2
D 52	0.72	0.55	0.48	0.43	0.42	0.39	0.46	2
D 53	0.32	0.29	0.25	0.22	0.16	0.23	0.22	2
D 54	0.59	0.20	0.19	0.14	0.12	0.22	0.17	2
D 55	0.13	0.09	0.10	0.07	0.04	0.11	0.08	2
D 56	0.34	0.47	0.44	0.37	0.37	0.41	0.42	2
D 57	0.48	0.16	0.19	0.14	0.12	0.21	0.17	2
D 58	0.45	0.34	0.31	0.28	0.19	0.32	0.29	2
D 59	0.74	0.37	0.40	0.38	0.37	0.28	0.38	2
D 60	0.35	0.40	0.36	0.30	0.28	0.34	0.34	2
D 61	0.76	0.00	0.00	-0.00	-0.04	-0.00	-0.02	2
D 62	0.30	0.30	0.32	0.29	0.22	0.31	0.30	2
D 63	0.82	0.45	0.49	0.51	0.39	0.35	0.47	2
D 64	0.11	-0.01	0.03	0.01	-0.02	0.04	0.01	2
D 65	0.74	0.51	0.49	0.47	0.42	0.37	0.47	2
D 66	0.52	0.75	0.64	0.62	0.54	0.53	0.63	2
D 67	0.38	0.19	0.20	0.19	0.14	0.14	0.18	2
D 68	0.77	0.62	0.60	0.60	0.56	0.42	0.59	2
D 69	0.41	0.48	0.40	0.40	0.28	0.33	0.38	2
D 70	0.75	0.38	0.36	0.36	0.27	0.29	0.34	2
D 71	0.61	0.46	0.40	0.35	0.32	0.34	0.37	2
D 72	0.73	0.56	0.53	0.53	0.43	0.40	0.51	2
D 73	0.11	-0.21	-0.28	-0.30	-0.26	-0.23	-0.29	2
D 74	0.56	0.33	0.30	0.27	0.28	0.18	0.28	2
D 75	0.65	0.34	0.30	0.30	0.25	0.19	0.28	2
D 76	0.21	0.10	0.14	0.13	0.14	0.04	0.12	2
D 77	0.41	0.16	0.12	0.08	0.12	0.05	0.09	2
D 78	0.65	0.61	0.51	0.43	0.53	0.39	0.49	2
D 79	0.68	0.51	0.44	0.35	0.46	0.35	0.42	2
D 80	0.62	0.60	0.51	0.44	0.52	0.38	0.49	2
D 81	0.75	0.35	0.36	0.28	0.38	0.27	0.34	2
D 82	0.75	0.48	0.48	0.44	0.47	0.33	0.46	2
D 83	0.60	0.35	0.28	0.24	0.29	0.15	0.26	2
D 84	0.68	0.07	0.09	0.09	0.10	-0.03	0.07	2
D 85	0.49	-0.27	-0.24	-0.30	-0.23	-0.17	-0.27	2
D 86	0.61	0.38	0.34	0.27	0.30	0.30	0.32	2
D 87	0.62	0.02	0.06	-0.03	0.11	0.04	0.04	2
D 88	0.61	0.08	0.12	0.04	0.17	0.05	0.09	2
D 89	0.37	0.30	0.30	0.19	0.38	0.22	0.28	2
D 90	0.56	0.37	0.36	0.27	0.39	0.28	0.34	2
D 91	0.63	0.38	0.38	0.28	0.42	0.28	0.36	2
D 92	0.64	0.49	0.45	0.37	0.49	0.32	0.43	2
D 93	0.29	0.24	0.29	0.20	0.30	0.24	0.27	2
D 94	0.35	0.10	0.16	0.09	0.15	0.15	0.14	2
D 95	0.65	0.20	0.20	0.18	0.13	0.17	0.18	2
D 96	0.80	0.45	0.50	0.43	0.49	0.39	0.48	2
D 97	0.38	0.28	0.29	0.26	0.23	0.24	0.27	2
D 98	0.94	0.15	0.32	0.29	0.32	0.19	0.30	2
D 99	0.36	-0.22	-0.17	-0.16	-0.15	-0.23	-0.19	2
D100	0.80	0.37	0.41	0.36	0.37	0.33	0.39	2
D101	0.62	0.29	0.26	0.17	0.23	0.25	0.23	3
D102	0.41	0.29	0.32	0.23	0.28	0.31	0.30	3

D103	0.41	-0.20	-0.15	-0.18	-0.18	-0.12	-0.18	3
D104	0.37	0.37	0.32	0.26	0.28	0.29	0.30	3
D105	0.33	0.20	0.22	0.16	0.17	0.22	0.20	3
D106	0.65	0.26	0.26	0.20	0.18	0.29	0.24	3
D107	0.40	0.19	0.20	0.16	0.10	0.22	0.17	3
D108	0.52	0.51	0.47	0.43	0.39	0.39	0.45	3
D109	0.44	0.51	0.43	0.33	0.38	0.42	0.41	3
D110	0.32	0.49	0.45	0.40	0.34	0.44	0.43	3
D111	0.44	0.65	0.57	0.52	0.48	0.50	0.55	3
D112	0.59	0.52	0.42	0.35	0.37	0.37	0.40	3
D113	0.26	0.37	0.36	0.32	0.29	0.32	0.34	3
D114	0.48	0.27	0.28	0.22	0.27	0.23	0.26	3
D115	0.46	0.43	0.39	0.34	0.32	0.33	0.37	3
D116	0.15	0.03	0.05	0.04	-0.02	0.07	0.03	3
D117	0.35	0.38	0.38	0.35	0.28	0.36	0.36	3
D118	0.58	0.38	0.33	0.28	0.29	0.26	0.31	3
D119	0.83	0.20	0.26	0.22	0.25	0.17	0.24	3
D120	0.68	0.45	0.43	0.38	0.40	0.34	0.42	3
D121	0.52	0.40	0.38	0.32	0.33	0.35	0.36	3
D122	0.51	0.43	0.39	0.33	0.32	0.38	0.37	3
D123	0.42	0.34	0.31	0.25	0.25	0.30	0.29	3
D124	0.35	0.02	0.01	-0.01	-0.00	-0.02	-0.01	3
D125	0.23	0.19	0.21	0.13	0.13	0.28	0.19	3
D126	0.45	0.37	0.35	0.31	0.27	0.32	0.33	3
D127	0.59	0.27	0.25	0.19	0.19	0.24	0.23	3
D128	0.62	0.54	0.49	0.45	0.43	0.37	0.47	3
D129	0.32	0.27	0.24	0.16	0.25	0.20	0.22	3
D130	0.17	0.02	0.03	0.03	-0.04	0.03	0.01	3
D131	0.63	0.09	0.14	0.06	0.10	0.18	0.12	3
D132	0.72	0.54	0.51	0.45	0.46	0.42	0.49	3
D133	0.24	0.24	0.24	0.20	0.22	0.19	0.22	3
D134	0.20	0.02	0.00	-0.08	0.01	0.05	-0.02	3
D135	0.33	0.49	0.47	0.38	0.41	0.45	0.45	3
D136	0.17	0.20	0.23	0.21	0.18	0.20	0.22	3
D137	0.19	0.11	0.12	0.13	0.05	0.10	0.10	3
D138	0.25	0.28	0.37	0.32	0.28	0.37	0.35	3
D139	0.18	0.08	0.13	0.09	0.10	0.12	0.11	3
D140	0.20	0.03	0.04	0.01	0.02	0.03	0.02	3

LISTED DISCRIMINATION INDICES ARE PROP(GROUP 3) - PROP(GROUP 1)
LISTED CORRELATIONS ARE POINT-BISERIALS BETWEEN ITEM SCORE
(0 OR 1) AND TOTAL SCORE.
SUB-SCALE CORRELATIONS ARE (UNBIASED) POINT-BISERIALS
BETWEEN ITEM SCORE AND SUB-SCALE SCORE.

TABLE OF INTERCORRELATIONS

2	0.782	1.000	
3	0.706	0.703	1.000
	1	2	3

NUMBER OF CANDIDATES = 266

TOTAL SCORE MEAN =	75.53	STANDARD DEVIATION =	20.16
SUB 1 SCORE MEAN =	31.32	STANDARD DEVIATION =	8.84
SUB 2 SCORE MEAN =	27.61	STANDARD DEVIATION =	7.26
SUB 3 SCORE MEAN =	16.60	STANDARD DEVIATION =	6.09

INTERNAL CONSISTENCY RELIABILITY (KR20) =		0.937
SUBTEST 1	" " =	0.894
SUBTEST 2	" " =	0.821
SUBTEST 3	" " =	0.790

Appendix IV
Revised Pilot Tests

IV. Revised pilot tests

The following items were selected on the basis of discrimination and content criteria to be used in the second pilot test:

1. Second Pilot A:

   a. Part 1: 1 2 3 4 5 6 8 10 12 13 14 15 16 17 19 20 28 29 32 33 34 36 40 41 42 44 45 47 49 50 [all from Form C: see Appendix II]

   b. Part 2: 51 53 54 55 56 57 58 59 60 61 62 63 65 67 71 72 73 78 79 80 81 82 83 84 95 96 97 98 99 100 [all from Form C]

   c. Part 3: 101–108 109–116 117–124 125–132 132–140 [all from Form C]

2. Second Pilot B:

   a. Part 1: 3 4 5 8 12 13 17 18 21 24 26 27 28 29 30 32 33 34 35 36 37 39 40 41 43 44 45 47 48 49 [all from Form D: see Appendix II]

   b. Part 2: 51 52 56 58 59 60 62 63 65 66 68 69 70 71 72 78 79 80 81 82 83 84 95 96 97 98 99 100 [from Form D] 64 68 [from Form C]

   c. Part 3: 109–116 117–124 125–132 [from Form D] 125–132 [from Form A] 117–124 [from Form B]

SUMMARY OF RESULTS OF SECOND PILOT

	Part 1	Part 2	Part 3	Part 4	Total
Test Length	30	30	40	30	
Reliability (A)	.90	.87	.81		.95
Reliability (B)	.87	.85	.61		.92
Mean (A) (n=603)	17.7	19.2	19.5	12.7	56.4
S.D. (A)	7.0	6.3	6.4	3.1	18.4
Mean (B) (n=689)	18.3	18.8	13.3	14.7	50.4
S.D. (B)	6.3	5.9	4.3	4.0	14.9
Correlations (A)					
Part 1	1.00	.83	.81	.25	.95

Part 2		1.00	.78	.26	.93
Part 3			1.00	.34	.92
Part 4				1.00	.38
Correlations (B)					
Part 1	1.00	.83	.66	.21	.94
Part 2		1.00	.67	.29	.93
Part 3			1.00	.41	.83
Part 4				1.00	.38
Criterion Correlations (A)					
SMP 1119 (n=135)	.64	.61	.67	.19	.69
SMP 322 (n=500)	.85	.77	.72	.13	.86
Criterion Correlations (B)					
SMP 1119 (n=144)	.68	.61	.65	.19	.71
SMP 322 (n=567)	.82	.75	.56	.36	.83